Institut d'études politiques de Paris
ÉCOLE DOCTORALE DE SCIENCES PO
Programme doctoral d économie
Département d'économie
Doctorat en sciences économiques

Essays on skills and education

Oliver CASSAGNEAU-FRANCIS

Thesis supervised by
Ghazala AZMAT, Professeur des Universités, IEP de Paris
Defended on 7 June 2022

Jury:
Emla FITZSIMMONS, Professor of Economics, UCL (reviewer)
Eric FRENCH, Professor of Economics, University of Cambridge (reviewer)
Clement DE CHAISEMARTIN, Professeur des Universités, IEP de Paris
Robert GARY-BOBO, Professeur des Universités, Université Paris 1
Jean-Marc ROBIN, Professeur des Universités, IEP de Paris

For my brother, Oscar.
I am proud to become the second Dr. Cassagneau-Francis—and I know you wouldn't have let me forget that you were the first. I miss you always.

Acknowledgements

First I would like to thank my advisors, Ghazala and Jean-Marc, who have been an invaluable source of support and encouragement throughout my PhD. Ghazala, although having you as my advisor was slightly serendipitous (given you joined Sciences Po after I started my PhD), you and Jean-Marc complemented each other perfectly, and your guidance and advice have contributed immeasurably to my development as an economist. Jean-Marc, I have looked up to you (both literally and figuratively!) since taking your econometrics class seven(!) years ago, and it has been a pleasure to be your student, coauthor and TA.

Next I would like to thank my coauthors, Julie and Robert who were there at the beginning of my PhD, and it appears our collaboration will continue after I complete it — a silver lining to our recent disappointment!

I would also like to thank all of my teachers, both in and outside of economics, who have instilled in me an interest in (and love of) education. A special mention goes to those for whom I was fortunate enough to TA: Eduardo Perez, Jean-Marc Robin, Stéphane Carcillo, Thomas Chaney and Basile Grassi. I believe teaching is a vital (and enjoyable) part of academia and I have benefited immensely from working with you.

Despite a number of unplanned interruptions to our time in the *combles*, the support and community provided by my colleagues at Sciences Po — many of whom were also my former TAs, classmates, or students — has been vital. I am grateful to the whole Sciences Po community — from the administrative staff whose tireless work ensures everything runs smoothly, to the students and faculty whose comments and investment have greatly improved my work. I would like to specially thank the professors who organised and attended the various incarnations of reading groups: Guy (whose presence is a reassuring constant), Pierre, Pierre-Philippe and Moshe. And my former classmates Pauline, Gustave, and Samuel for organising the PhD seminar series.

A special mention is also reserved for the various footballers who I have played with throughout my time at Sciences Po. The games amongst the PhD students have had a distinctly Italian flavour, and Marco's organisational talents ensured it became a weekly necessity, and a welcome break from studying and research.

A number of those who I met during my studies in economics are now some of my closest friends. Max, always up for a beer, I still miss our weekly futsal-and-lunch dates.

Marcus and Andreas, with whom I both began my journey as an economist in Cambridge, and recently celebrated another milestone in Paris. Then, there are those whom I am in the midst of visiting in the US as I write this. Claire, Fabio, Hugo, Nicolo, and Théo: we've enjoyed so much together, from discussions over a beer at (the) Supercoin, to holidays by the sea and in the mountains. Caro, Pia and Quentin shared in many of those adventures despite having to suffer discussions of economics.

Then there are my friends from outside of economics, who have been a vital source of escape and support since before I began my PhD and will continue to be for many years. The "Rat Pack"— George, Henry, Jim, Pete and Sam — who I have known since my first year in Birmingham. We have lived together, cooked together, holidayed and partied together, and more recently it has been marriages (and births!) to celebrate. We, along with the other members of the "Book Club" who I have met since, have shared the best of times and our meetings are always highlights of my year. The Cambridge Crew: Mike, Rob and Ravi. Life in Cambridge was always the better for seeing you (and we'll always have Barcelona). And Brocks: your support and advice (having been through this) were invaluable—and living with you next year is going to be so much fun! I'm very happy that in moving back to England I will be closer to (almost) all of you.

Finally I thank my family. Daniel, I'm not sure what title to give you, but your support has really helped me over the past 7 years; from advice on academia, sharing your love of travel, to being the best "landlord" I could ever hope for!

My parents: I know you may not quite understand what I've been doing these past 5 (7!) years, your unconditional love and support has been unwavering. Your approach to life has made me the person I am today, and I am so lucky to have such amazing people as parents. Your (our) house in Cambridge will always be my home, and being able to come back and spend time away from the PhD life has been so important to me (even if recently that included the refrain "Finished that PhD yet, boy!". I can now happily say "Yes!").

And Laélia. I still remember vividly the first time we met, when I was behind a bar rather than a computer. It has been an incredible journey for both of us, and I could not (and would not) have done it without you. Your ambition and enthusiasm to be the best you can and make the most of everything is infectious. You are the most incredible person, and I am so lucky to have met you.

Contents

List of Figures

List of Tables

Introduction

The three chapters of this thesis study different aspects of skills and human capital across different contexts: the determinants (chapter 1) and returns (chapter 2) to higher education in England, and the returns to formal training in France (chapter 3).

The first chapter investigates the following question: what drives some young people to choose to attend university while others do not? A novel feature of this chapter is a focus on non-pecuniary factors: using detailed survey data from UK cohort studies I compare the expectations of young people who both attend and do not attend university about many different aspects of life at, and after, university. Non-pecuniary, and particularly non-financial, factors are shown to be much more important than wages in determining the higher education choices of young people.

The second chapter moves the focus of the analysis from expected returns from before attending university, to realised returns after university — from *ex ante* to *ex post.* Using the same cohort studies as in chapter 1, I study the wage returns to a university degree in the UK as a function of ability on entry to university. I develop a methodology that allows the estimation of both cognitive and non-cognitive prior abilities. I use these estimates to study how the returns to university vary across groups of individuals with different prior ability levels, allowing for interactions between the different components of ability.

In chapter three the focus switches from higher education to formal training, still keeping with the theme of skills and human capital investment. Using a novel methodology in the spirit of difference-in-differences, which specifically allows for unobserved heterogeneity, we exploit novel French data on training to estimate the wage returns to formal training. We find small estimates in the range of 1-3%, suggesting the larger estimates found in earlier studies may have failed to fully account for unobserved heterogeneity. In general the work in this thesis emphasizes the importance of higher education in explaining wage inequality, while also highlighting the multiple dimensions that young people consider when deciding whether to make this investment in their future.

Chapter 1 — The role of earnings and other factors in university attendance

Typically, economics has focused on the importance of the graduate wage premium as the key driver of university attendance, and on credit constraints as the main barrier to investment in human capital. However, recently researchers have highlighted the failure

of these purely pecuniary factors to fully explain observed educational and occupational choices (Cunha and Heckman, 2007; D'Haultfoeuille and Maurel, 2013; Arcidiacono et al., 2020). In this chapter, I shed new light on this key issue by comparing and quantifying the roles of earnings expectations and non-pecuniary factors in the decision to attend university in the UK. Recent work has made important advances to explore beyond wages in explaining educational and occupational choices, mainly in the US (Cunha and Heckman, 2007; Arcidiacono et al., 2020). These papers rely on a residual term to capture non-pecuniary factors—as they lack a direct measure—and find they play a major role in both educational and occupational decisions. Boneva and Rauh (2020) are an exception: they implement a survey of students in secondary education in the UK, eliciting expectations about both pecuniary and non-pecuniary factors.

My contribution builds upon this work, exploiting rich data on both observed outcomes and young people's expectations about non-pecuniary factors. I estimate a model of university choice using panel data from a representative sample, which contains young people's expectations about the future and their realised outcomes. I use my model to investigate the factors affecting university attendance in England, answering three key questions: (i) How important are expectations about earnings versus other factors for 16–18 year olds when deciding to go to university? (ii) What drives the educational attainment gap between advantaged and less-advantaged potential students? (iii) How has the importance of these factors in the decision changed between the 1980s and today?

I find an even larger role for non-earnings factors than in previous work, with non-pecuniary factors four times as important as earnings in the decision to attend university. This result emphasizes the range of costs and benefits that young people consider when making educational decisions. I also find that earnings expectations are similar across socio-economic groups, suggesting differences in other factors are almost entirely responsible for the observed gap in attainment. The current gap in attainment between those from advantaged and less-advantaged backgrounds is not driven by differences in (expected) earnings, nor by difficulties in obtaining funding. To address this socio-economic imbalance policymakers should focus on other aspects of university life, aspects that are easier to affect than earnings and cheaper than reducing tuition fees. I use detailed information on young people's expectations about life at, and after, university to decompose *other factors* into more meaningful, and more policy-relevant, categories. Separating expectations about debt and the monetary costs of attending university from the other factors, I find that financial factors do not play a major role in the decision. On the contrary, it appears that young people who are most concerned about the impact of student loan debt—and other monetary costs of attending university—are those who are most likely to attend. Finally, there appears to have been little change in the *ex ante* pecuniary returns to a university degree in recent decades, with the increase in attendance driven by improved expectations about the non-pecuniary factors associated with a university degree.

Chapter 2 — The wage returns to higher education by skills and ability

While the first chapter was concerned with the *ex ante* (expected) returns to university, the second estimates the *ex post* (realised) wage returns to a degree. This joins a large literature on the returns to education, from the (average) effects of an additional years schooling, to more precise analyses of the effects of passing key educational milestones—i.e. graduating high school or college. A key difficulty has been to separate the effects of (prior) ability from those of education, an issue made even more problematic by recent evidence on the large heterogeneity of returns to education. James Heckman and coauthors pioneered a methodology to address this difficulty, using (noisy) measurements of ability along with observed later outcomes to capture (and control for) *unobserved* ability and heterogeneity in returns to education. A different but related literature has highlighted the importance of both cognitive *and* non-cognitive skills for success—at school and in later life. Heckman and coauthors have extended their method to allow multiple components of ability, but as they rely on factor models these must enter wage or production functions linearly, prohibiting interaction between cognitive and non-cognitive skills.

I develop a novel framework to estimate the returns to human capital investments as a function of an individual's prior cognitive and non-cognitive abilities, allowing for interactions between different components of ability, and use my framework to estimate the wage returns to a university degree in England. I build upon recent work studying the formation of human capital (see Cunha et al. 2006 for a review), and incorporate insights from the long literature on the returns to education (see Card 2001 for a review). I then use my framework to answer two key questions: how does one separate the effects of (cognitive) ability from the effects of higher education? With the brightest students attending the best universities, are these universities adding a huge premium onto the wages these students can command; or would these high ability students have received higher wages even without a degree?

In answering these questions, a key contribution is methodological. I demonstrate the nonparametric identification of, and a parsimonius estimation strategy for, two key (and related) objects: (1) a young person's cognitive and non-cognitive abilities at the time they would enter university; (2) the (wage) returns to university degree as a function of these abilities on entry. Using recent advances in the identification of discrete mixture models (Bonhomme et al., 2017), my identification strategy requires fewer observations and permits a more flexible functional form than the current leading approaches in the literature (see for example Cunha et al., 2010). I achieve this in a tractable framework by assuming the distribution of human capital has discrete support, and those individuals with the same level of human capital are said to be of the same latent type. Nonparametric identification ensures I do not need to rely on any difficult-to-defend parametric or

functional form assumptions. Identification requires an additional (crude) measurement or an instrument, a variable that measures or affects investment in human capital (i.e. university attendance), but is independent of wages, conditional on human capital before the investment. My framework allows for non-linearities in the returns to higher education by cognitive and non-cognitive abilities that are not captured by the current leading approaches in the literature. I show these non-linearities are empirically important in an application to higher education in England.

I apply my framework to data from the British Cohort Study, a longitudinal dataset containing information on 16,000 people born in a single week in April 1970. The results suggest important returns to university for those with low human capital on entry: they "catch up" in wage terms with their non-university educated peers who possessed much higher levels of human capital before university. However, the wage returns to university are generally increasing in prior human capital, meaning that a university degree increases wage inequality among graduates. This is unsurprising when viewed in light of research on human capital formation in childhood, which finds that "skills beget skills, [and] learning begets learning" (Cunha et al., 2006, p. 799), i.e. pre-existing human capital increases the efficiency of further investments in human capital. I also find strong evidence of non-linearities in the returns to higher education, in particular in the interactions between cogntive and non-cognitive skills. These interactions are not possible to capture using the usual approaches to estimation which rely on an additive model for identification and estimation.

Chapter 3 — A nonparametric finite mixture approach to difference-in-difference estimation, with an application to on-the-job training and wages

with Robert GARY-BOBO, Julie PERNAUDET, and Jean-Marc ROBIN

The third chapter departs from the study of higher education to focus on another key investment in human capital: formal training. Training has long been a focus of policy-makers seeking to ease the transition of workers from industries disrupted by automation and trade. It also appears a natural substitute for tertiary education, potentially helping to ameliorate the effects of higher education on wage inequality highlighted in chapter 2. However, the evidence on the wage returns to higher education is mixed, with most authors finding small positive effects, though more recent analyses have found effects of up to 10%. However, take up of training is generally low, with 40% of workers training in France and only 20% in the Netherlands. Such large returns to training are at odds with the perceived impact of training. Could selection, unobserved heterogeneity, and endogenous treatment lead to such important biases that standard statistical evaluation methods (instrumental variables, selection models, etc.) are unable to provide reliable

estimates? Or is the perception of the effects just wrong? Our paper makes original contributions to this question, both empirically and methodologically.

We develop and apply a novel methodology to new linked employee-employer survey and administrative data to measure the wage returns to training in France. Our approach is in the spirit of difference-in-difference estimation, but we use a combination of economically motivated exclusion restrictions and discrete mixtures (to capture unobserved heterogeneity) to relax the common-trends assumption usually required in such analyses. We prove the non-parametric identification of our discrete-mixture model, and demonstrate a viable estimation strategy via the expectation-maximisation (EM) algorithm of Dempster et al. (1977).

Empirically, we find small average effects of training on wages of around 1% depending on our specification. Our framework allows the returns to vary across *types*,[1] and we find significant heterogeneity in the effects of training across these different types. For some types we estimate treatment effects of over 10%, while for others the effects of training on wages are slightly negative. These findings are important given the focus of governments around the world on training as a solution to labour market changes, especially those driven by technology. If the benefits of training are different across different workers, then misinformed policies towards worker training could lead to inefficiencies and even increases in wage inequality. Another direct relevance of our work for policy is the effectiveness of our policy variable in encouraging people to train: the provision of information on training opportunities to workers by their employers.

Bibliography

Arcidiacono, P., V. J. Hotz, A. Maurel, and T. Romano (2020): "Ex Ante Returns and Occupational Choice," *Journal of Political Economy*, 128, 4475–4522.

Boneva, T. and C. Rauh (2020): "Socio-Economic Gaps in University Enrollment: The Role of Perceived Pecuniary and Non-Pecuniary Returns," *HCEO Working Paper*, 83.

Bonhomme, S., K. Jochmans, and J.-M. Robin (2017): "Nonparametric estimation of non-exchangeable latent-variable models," *Journal of Econometrics*, 201, 237–248.

Card, D. (2001): "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69, 1127–1160, publisher: [Wiley, Econometric Society].

Cunha, F. and J. J. Heckman (2007): "Identifying and Estimating the Distributions of Ex Post and Ex Ante Returns to Schooling," *Labour Economics*, 14, 870–893.

[1] We use latent discrete "types" to capture unobserved heterogeneity.

Cunha, F., J. J. Heckman, L. Lochner, and D. V. Masterov (2006): “Interpreting the Evidence on Life Cycle Skill Formation,” in *Handbook of the Economics of Education*, Elsevier, vol. 1, 697–812.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010): “Estimating the Technology of Cognitive and Noncognitive Skill Formation,” *Econometrica*, 78, 883–931, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA6551.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977): “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the royal statistical society. Series B (methodological)*, 1–38.

D’Haultfoeuille, X. and A. Maurel (2013): “Inference on an extended Roy model, with an application to schooling decisions in France,” *Journal of Econometrics*, 174, 95–106.

Chapter 1

Earnings, financial and non-pecuniary factors in university attendance

Abstract

Why do some people choose to attend university, and enjoy state-subsidised benefits, while others do not? We shed new light on this key issue by comparing and quantifying the roles of earnings and non-pecuniary factors in the educational decisions of young people in the UK, exploiting information on young people's beliefs about the advantages and disadvantages of university. We also investigate changes in these factors over time, and their implications for social mobility. We specify a model of educational choice, explicitly including expectations about earnings, financial, and non-pecuniary factors. Our estimation strategy exploits panel survey data on young people's expectations about key outcomes both at, and after, university, linked to their realised outcomes. Income maximisation, despite its prevalent role in the literature, is only part of the story: other factors are at least as important as earnings in determining whether someone goes to university. Non-pecuniary factors also play an important role both the SES-gap in educational attainment, and the huge growth in degree attainment between the 1980s and 2010s.

1.1 Introduction

University graduates enjoy higher wages, better health, and are more likely to report feeling happy with their lives than their less-educated peers (Heckman et al., 2018; Oreopoulos and Petronijevic, 2013; Oreopoulos and Salvanes, 2011). These benefits are often heavily subsidised, with governments in developed countries spending around 1% of their countries' GDP on higher education (OECD, 2018). Therefore, understanding the determinants of young people's decision to attend university is key, not only to better understand the direct effects of educational policies, but also for wider issues such as inequality and to identify the beneficiaries of public spending. Traditionally, economists have focused on higher wages, pecuniary costs, and financial frictions to explain this decision—a narrative of comparative advantage and credit constraints. However, more recent work suggests this narrative is missing an important part, as "[t]he evidence against strict income maximization is overwhelming" (Heckman et al., 2006, p. 436).

A growing literature has attempted estimating these "psychic costs", as the non-earnings factors are commonly termed. Authors have typically used a residual term to capture these factors, relying upon data containing information on family background, earnings, and educational choices (Cunha and Heckman, 2007). However, some of these same authors have highlighted issues with relying on a residual catch-all term, with Heckman et al. (2006, p. 436) remarking that "explanations based on psychic costs are intrinsically unsatisfactory [as o]ne can rationalize any economic choice data by an appeal to psychic costs."

In this paper, we study the role of both pecuniary and non-pecuniary factors (or "psychic costs") in the decision to attend university, with the aim of exploiting data which combines young people's subjective beliefs about the (pecuniary and non-pecuniary) aspects of university, with information on their later outcomes and educational choices. Our main analysis addresses the following question: what is the relative importance of wages versus other non-pecuniary factors in the decision to attend university? We then extend our analysis to study heterogeneity in educational decisions across different socio-economic groups and over time.

Our data is from a longitudinal study of young people in the UK, which follows a representative sample of students born in 1989 or 1990. The cohort members were surveyed annually between the ages of 13 and 18 — a period in which they were in compulsory education (up to 16), and then either transitioning to work, or on to further and higher education. They were contacted and surveyed again at age 25. The dataset contains information on: (i) young people's beliefs about (the advantages and disadvantages of) university obtained prior to their decision to attend; (ii) their educational and career choices; (iii) and their later wages.

Given that one of our chief aims is to quantify the non-pecuniary factors in the decision

to attend university, these subjective beliefs about university, which include many non-pecuniary aspects, are central to our analysis. Examples of these non-pecuniary factors are: the effort required to gain a place at university; aspects of life at university (social life, studying, leaving home, stress, etc); and aspects of life after university (access to better jobs, graduate "identity", debt). These beliefs are recorded in the form of open-ended questions about the subjective advantages and disadvantages of attending university, and were obtained from a representative sample of young people. As far as we know, we are the first to analyse educational decisions using data containing detailed information, including elicited beliefs about non-pecuniary factors, from a representative sample including data on realised outcomes after university. Previous work has relied on small, selected samples (Boneva and Rauh, 2020), or did not have access to information on young people's beliefs (D'Haultfoeuille and Maurel, 2013).

To guide our empirical analysis, we specify a parsimonious model of educational choice in the spirit of Roy (1951) explicitly including both earnings and other (chiefly non-pecuniary) factors. The structure of the model allows us to quantify and compare the relative contributions of different factors in the decision to attend university, exploiting choice data, subjective beliefs about life at and after university, and realised earnings. We map observed (realised) earnings into potential (expected) earnings as the mean of realised earnings at age 25 conditional on a set of observed characteristics at age 16. The model is then straightforward to estimate using standard techniques from the discrete-choice literature (Mcfadden, 2001). Having estimated our model, we are able to combine estimated preferences with (observed and estimated) expectations to obtain distributions of the relative contributions of earnings and other factors to the decision to attend university. These distributions are: (i) students' expected earnings premium, and; (ii) students' (observed) expected "other factors premium", from attending university. We rescale both distributions of premiums so that they are expressed as a percentage change in wages, allowing a direct comparison.

Our results highlight the important role for non-pecuniary factors as a determinant of higher education attendance. Although the distributions of earnings and of other factors share similar shapes and locations — bell-shaped, with slightly positive means — the dispersion of the other factors distribution is twice as high. This much larger dispersion, along with the similar shapes and locations of the distributions near zero, suggest that the chief determinant of whether or not someone decides to go to university is their expectations about factors other than their future earnings. To underline this conclusion, we study the effects of varying values of these factors, performing the same counterfactual exercise as D'Haultfoeuille and Maurel (2013). This involves fixing the values of young people's pecuniary and non-pecuniary factors at certain percentiles of the distribution, and recalculating the proportion who get a university degree under these counterfactual values. Assigning everyone in the sample an "other factors premium" equal

to the 10th percentile results in 35% obtain a university degree; assigning them values equal to the 90th percentile results in 86% graduating — a change of over 50pp. Repeating the same exercise with earnings (assigning different values of the earnings premium to everyone) results in 50% (10th percentile) and 73% (90th percentile) of people attending and graduating university, a much smaller variation.

Next, we split the sample into three groups by socio-economic status (SES) measured by parental earnings at sixteen,[1] to investigate the role of earnings and other factors in the socio-economic gap in university attainment. We recalculate the distributions of earnings and other factors premiums for each of the three SES groups. The distribution of the expected graduate earnings premium is remarkably stable across the three groups, with means ranging from 7 log-points (low SES) to 9 log-points (high SES), and dispersion slightly decreasing with parental income. For other factors, there is much more variation across SES: the low-SES mean is 3 log-points, while the high-SES mean is 15 log-points. The socio-economic gap in university attainment is almost entirely driven by factors other than earnings.

Finally, we re-estimate the model on data from an earlier cohort born in 1970. Comparing the distributions of earnings and other factors from the earlier with the later cohort allows us to assess their role in the huge growth in higher education seen over this period. The distribution of the expected graduate-wage premium remained quite stable over this period, with its mean and dispersion only decreasing slightly. The distrubution of the other factors premium, however, changed drastically, shifting right so that the strongly negative mean of the 1970 cohort became slightly positive for the 1990 cohort. The variance of other factors premium distribution also increased slightly. Taken together these results suggest the increase in degree attainment—which went from 29.9% in the 1970 sample, to 62.2% in the 1990—was entirely driven by changes in the other factors premium.

1.1.1 Related literature

This paper joins a long tradition of studying educational and occupational decisions in economics and social science. Arguably this tradition in economics begain with Roy's seminal 1951 paper on occupational choice. Roy models (and their extensions) have since been applied to educational choice, a literature which includes an important series of papers by Cunha, Heckman and coauthors (Cunha et al., 2004; Cunha and Heckman, 2007; Heckman et al., 2006). These and more recent papers highlight the importance of non-pecuniary factors (often called "psychic costs") in explaining educational choices, both at the intensive (e.g. major choices in Wiswall and Zafar (2015) and institutional choices in Delavande and Zafar (2019)) and extensive margins (D'Haultfoeuille and Maurel, 2013;

[1] These correspond to the bottom quintile of parental earnings in the sample (low SES), the middle-three quintiles (middle SES), and the top quintile (high SES).

Boneva and Rauh, 2020). Arcidiacono et al. (2020) show the importance of non-pecuniary factors in occupational choice.

Our contributions to this literature are the following. We exploit information on realised earnings and choices, linked to young people's subjective beliefs about the advantages and disadvantages of attending university, including about the non-pecuniary aspects. A growing literature studies young people's choices by eliciting expecations about future earnings from students, but in general these do not elicit expectations about non-pecuniary factors (Manski, 1993; Dominitz and Manski, 1996; Arcidiacono et al., 2012, 2020).[2] For much of this important work realised outcomes are not (yet) available.[3] In addition, most of the prior work using elicited expectations has often used smaller, selected samples, either from a single US college (Arcidiacono et al., 2020) or self-selected survey respondents (Boneva and Rauh, 2020). Our data is from a large, representative sample.

There is also a growing recent literature on the differences across socio-economic groups in education attainment and choices, work to which this paper is closely related. The persistence of the educational attainment gap between more- and less-advantaged students, even during a period of huge expansion in higher education, is documented by Blanden and Machin (2004). More recently, differences across social groups in terms of subject and institution have been highlighted as a key driver of differences in labour market outcomes (Britton et al., 2021). The drivers of these differences are also beginning to be explored. Boneva and Rauh (2020) study the role of students' beliefs about pecuniary and non-pecuniary outcomes in the decision to attend university in England. Anders (2012) and Anders and Micklewright (2015) investigate how young people's expectations about applying to university evolve differently between the ages of 14 and 17 according to their socio-economic group, using the same cohort study that we analyse in this paper. We contribute to this literature by comparing and quantifying earnings and other factors in the decision to attend university across socio-economic groups.

Finally, through our analysis of the evolution of factors in the decision to attend university over two decades, our work is related to the literature studying the recent growth in university attendance in the UK. Higher education in England has seen substantial changes in recent decades, undergoing substantial growth and an overhaul of its funding system. Growth in attainment has been much steeper in the UK than in the US. The proportion of UK (US) BAs in a given cohort at age 30 increased from less than 10% (25%) for those born in 1950, to nearly 40% (35%) for those born in 1985 (Blundell et al., 2021). Alongside this rapid growth in attainment, the graduate wage premium has remained flat in the UK, while it has been steadily increasing in the US (Blundell et al.,

[2] Boneva and Rauh (2020) is a notable exception.

[3] Outcomes are beginning to become available for some surveys which elicited expectations (see for example Arcidiacono et al. (2020) and Gong et al. (2019)).

2021). However, this recent growth in higher education in the UK did not occur equally across socio-economic groups, with the children of richer parents disproportionately benefitting from the expansion (Blanden and Machin, 2004). Walker and Zhu (2008) study the impact of this expansion on the graduate wage premium, finding no change for men, and a slight increase for women. Such rapid growth in higher education, over a period of increased fees and stagnant returns, raises questions about what drove so many more people to attend university — questions we shed new light upon in this paper.

Outline. The rest of the paper is organised as follows. Section 1.2 describes the data we use in this paper, and presents some initial analysis. Section 3.2 describes the model we estimate to obtain our main results, with our estimation stratgy in section 2.3. Our main results are in section 2.5, followed by analysis by socio-economic group (section 1.4.2) and over time (section 1.4.3). Section 1.5 concludes.

1.2 Context and data

The data used in this paper come from *Next Steps*, a British cohort study which follows a representative sample of 15,770 people born in England in 1989 or 1990 (IOE Centre for Longitudinal Studies, 2018). These young people were able to leave school at age 16, in 2006, and those who went on to higher education would have entered university at age 18 or 19, in 2008 or 2009. They made the decision to apply to university at age 16 or 17, after deciding whether to stay in full-time education after 16. These students would face tuition fees of around 3,000 GBP per year, though there are extensive government-provided loans and grants available to cover both tuition fees and living costs. A detailed discussion of the application process, and the UK system of tuition fees, loans and grants is in appendix 1.A. Typical university degrees in the UK are 3-year bachelor's degrees, with a few subjects offering longer courses as standard.[4] Therefore, the majority of young people who choose to attend university will have graduated by the time they are 25.

Two important periods for this paper are just before young people apply to university (age 16 or 17), and when the majority have entered the labour market including those who attend university (age 25). We have data on the *Next Steps* cohort members at both these key stages. Data collection involved annual face-to-face interviews between 2004 and 2010 (waves 1–7), plus a further round of interviews in 2015 (wave 8).[5] In our analysis we use information on: schooling, family background, and subjective beliefs about university and the future at age 16 (before applying to university); and on earnings and qualifications at age twenty-five (after entry to the labour market). A selection of these variables are

[4]Engineering is often a 4-year combined bachelor's and master's degree, while an undergraduate medicine degree takes at least 5 years.

[5]The study is ongoing and the cohort members will be interviewed again in 2021, with plans to make the data available by 2023.

Table 1.1: Description of selected variables

Variable	Description
Earnings	Usual weekly earnings in GBP reported by the CM if employed at age 25 (wave 8).
Education	An indicator variable for whether the CM reports having an UG degree at age 25 (wave 8).
# of A-levels	The number of A-levels the CM reported taking at age 16 / 17 (wave 4).
Parental income	Total annual parental income when CMs were 16/17 (wave 4). The data is recorded in 12 bins.
Subjective beliefs	Harmonised, open-ended responses to questions about the advantages and disadvantages of attending university (wave 4).

described in table 2.1. Importantly, we have a direct measure of students' beliefs about the future. We supplement the data from Next Steps with data from the earlier British Cohort Study (BCS) to analyse changes in factors across time. The BCS is a similar study to Next Steps, following nearly 17,000 people born in the UK in a single week in April 1970.

Data description. Table 1.2 presents summary statistics from waves 4 (CMs aged 16) and 8 (CMs aged 25) of Next Steps, for all cohort members in our sample, and then split by whether they hold a degree at 25. Only those with a minimum of 5 GCSEs at A*-C or equivalent were asked about their subjective beliefs about university, information vital to the analysis in this paper.[6] The young people not asked about their expectations are not included in the analysis. We also drop young people for whom we do not observe a wage at age 25 (and those who reported wages above the 99th or below the 1st percentiles), who did not respond in either wave 4 or wave 8, who did not answer question about their perception of their ability, or who did not provide information on their qualifications at age 25. Although dropping students without 5 GCSEs at A*-C or equivalent reduces the size of our sample significantly, those omitted are likely students who would have found it very difficult to attend university. They are an important group to study, but their omission is not fatal to the current analysis.

1.2.1 Subjective beliefs about university

Information on students' beliefs about their future potential life at and after university is a key feature of this paper. These subjective beliefs were collected as open-ended responses to two questions, one about the advantages and one about the disadvantages of

[6]These are referred to as "high-achieving" students in the survey documentation (IOE Centre for Longitudinal Studies, 2008). Blundell et al. (2021) consider grade C at GCSE as the UK equivalent to high-school graduation in the US.

Table 1.2: Descriptive statistics

	All	$D=0$	$D=1$
N	3,469	1,311	2,158
Female	0.57	0.56	0.57
Ethnicity			
White	0.73	0.81	0.68
South Asian	0.16	0.10	0.20
Black	0.04	0.03	0.05
Other	0.07	0.06	0.07
Main parent's SOC			
SOC 1–3	0.42	0.35	0.46
SOC 4–5	0.22	0.24	0.21
SOC 6–9	0.36	0.41	0.33
Self-assessed ability†	0.25	0.08	0.34
# A-levels	3.74	3.43	3.92
Degree	0.63	0.00	1.00
Earnings at 25 (weekly in GBP)			
Mean	451.90	416.58	473.01
Std. dev.	197.00	192.84	196.44

Notes: †A composite measure combining students' response to asked a series of questions about their perceived abilities in maths, science, and english. All information except earnings and qualification was recorded at age 16 or 17, in waves 4 and 5.

university.[7] The CMs could mention as many or as few advantages (disadvantages) as they wished. The interviewer noted down their interviewee's responses, and similar responses across individuals were then grouped into the harmonised responses we use in our analysis. These harmonised responses are listed in tables 1.3 (advantages) and 1.4 (disadvantages), grouped into the following categories: career (non-pecuniary); earnings; financial / debt; social life / environment; education; and time. These tables also display the proportion of young people who mentioned each response, overall and separately for graduates and non-graduates. The final column is the difference in proportion of graduates ($D=1$) and non-graduates ($D=0$) who mentioned each response, expressed in percentage points (*pp*).

Figure 1.1 also shows the overall numbers of young people who mentioned each recorded advantage and disadvantage of attending university, ordered by number of mentions, rather than in categories. Focusing first on the reported advantages in figure 1.1a, access to "better opportunities" and to "better jobs" were the two most common advantages of a university degree mentioned by respondents. In close third was "more qualifications", with getting a "well-paid job" in fourth place. An enjoyable "social life" rounds out the top five most popular advantages, with "learning", "personal

[7]The exact wording of the question(s) was: "What do you think the advantages (disadvantages), if any, might be for someone of going to university to study for a degree?" (IOE Centre for Longitudinal Studies, 2008).

development", and "gain life skills" also popular responses. Although some of the responses are arguably linked to higher pay, there are many that are not, for example "social life" and "personal development". In addition, the presence of "well-paid job" as a specific response suggests other career-related responses are capturing broader notions than pay alone.[8] Turning to the disadvantages in figure 1.1b, the three responses mentioned most often are all financial concerns: "get into debt", "costs (general)", and "too expensive" — concerns which arguably still "pecuniary". However, many of the disadvantages mentioned reflect fully non-pecuniary aspects of a person's career ("no job guarantee"), or life at university ("heavy workload", "leave home"). Together these responses provide information on students' beliefs about the pecuniary, financial, and non-pecuniary aspects of attending university.

Returning to tables 1.3 and 1.4 we can take a higher level view, with the advantages and disadvantages displayed in broad categories. We can also see which categories were the most mentioned by young people. Focusing first on the advantages in table 1.3, the two most mentioned advantages are related to the (non-pecuniary) aspects of a person's future career. This is the most mentioned category, with over 66% of the sample mentioning something to do with their career as an advantage. Education is the next most popular category, followed by earnings, personal development, and then social life, which is still mentioned by nearly 17% of our sample.

1.2.2 Beliefs and university attendance

We also compare the answers of young people who go on to attend university ("graduates") and those who do not ("non-graduates"). The third and fourth columns of tables 1.3 and 1.4 show the proportion of respondents who mentioned each advantage or disadvantage of university separately for graduates ($D = 1$) and non-graduates ($D = 0$). The final column in these tables shows the difference in proportion between graduates and non-graduates mentioning each aspect of university. Generally the advantages of university were more likely to be mentioned by graduates (signalled by a positive difference in the final column of table 1.3), with all but one of the top-8 advantages being mentioned by a higher proportion of graduates.

The disadvantages are more balanced in terms of who is more likely to mention them, with the majority of differences in the final column of table 1.4 smaller than $2pp$ in magnitude. Still, somewhat surprisingly, the top 3 disadvantages are all more likely to be mentioned by graduates (and are all related to financial aspects of attending university). Our initial analysis suggests those who go on to university are more likely to believe it will be beneficial for their career (and earnings), for their personal development, and for their

[8]While it is difficult to know exactly what respondents meant by "better opportunities" and "better jobs", the presence of another (harmonised) response specifically referring to "well-paid job" suggests these might refer to a broader notion of quality than is captured by pay alone.

Table 1.3: Students' subjective beliefs about university (advantages)

	Prop. mentioning (%)			
Response (harmonised)	All	$D=0$	$D=1$	Diff. (*pp*)
Career (non-pecuniary)	66.27	58.52	70.91	12.39
Will lead to a good / better job (than would otherwise get)	32.37	29.01	34.32	5.31
Gives someone better opportunities in life	33.12	27.66	36.36	8.71
Is essential for the career they want to go into	3.32	3.80	3.02	-0.78
Shows that you have certain skills	1.82	1.86	1.78	-0.09
To delay entering work / more time to decide on a career	0.63	0.54	0.67	0.13
Earnings				
Will lead to a well paid job	21.33	19.29	22.51	3.22
Social life / environment	16.18	13.95	17.51	3.56
The social life / lifestyle / meeting new people / it's fun	15.22	13.23	16.38	3.16
To leave home / get away from the area	2.02	1.76	2.17	0.42
Education	34.01	35.31	33.23	-2.08
To carry on learning / I am good at / interested in my chosen subject	8.65	7.07	9.57	2.50
Get more / better / higher qualifications	26.38	28.77	24.90	-3.87
Personal development	17.68	13.52	20.17	6.65
Makes someone independent / maturity / personal development	8.76	6.04	10.36	4.33
Gives you more confidence	0.95	0.48	1.22	0.75
People will respect me more	0.35	0.34	0.35	0.01
Leads to a better life / good life (general)	2.13	1.89	2.27	0.38
Prepare you for life / gain life skills	7.52	6.33	8.25	1.92

Notes: Students with at least 5 GCSEs at A*–C or equivalent were asked these questions ($N = 3{,}469$). Where there are values in the right hand columns for the categories (in bold), these are the proportion of young people who mention at least one of the advantages in that category.

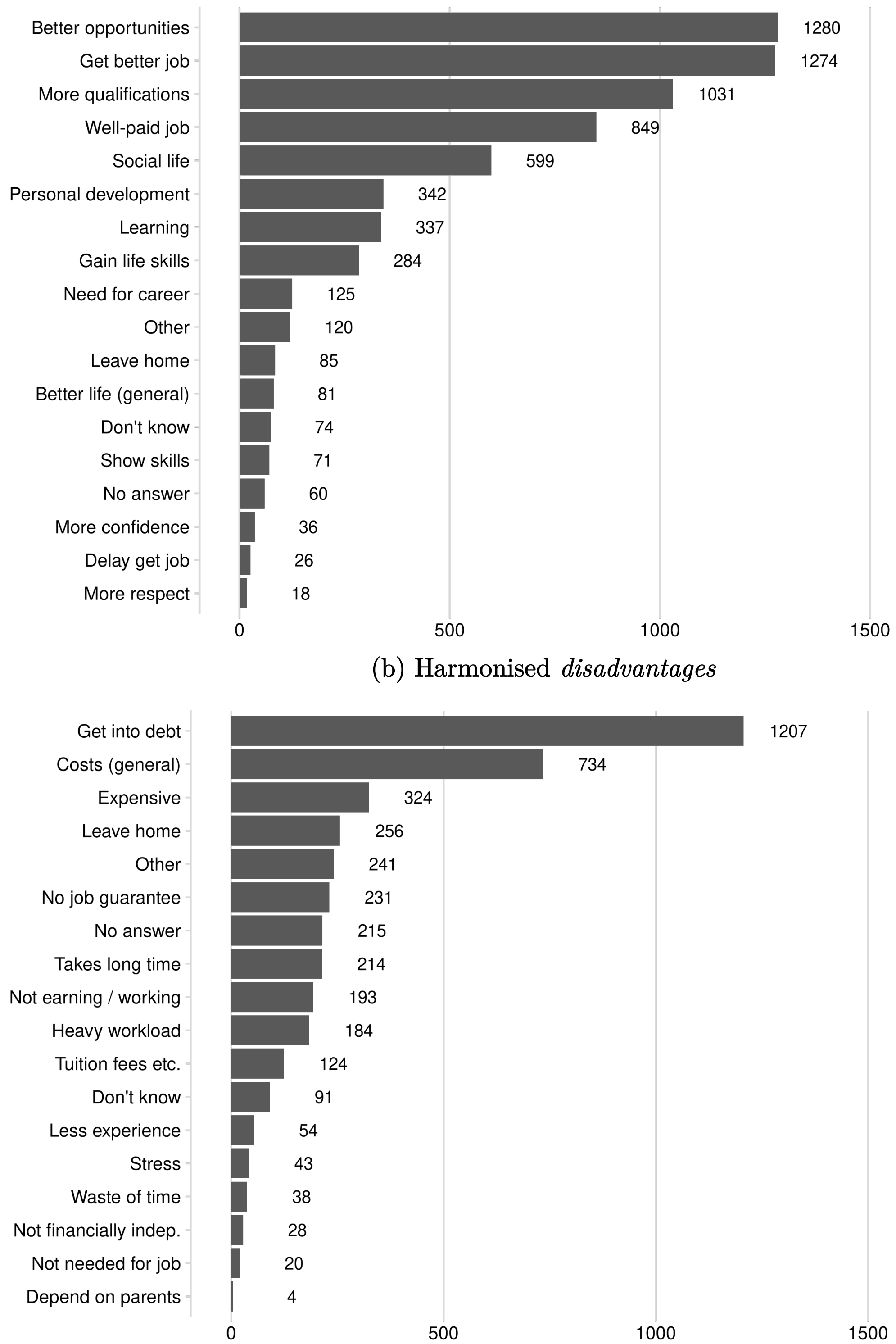
Table 1.4: Students' subjective beliefs about university (disadvantages)

	Prop. mentioning (%)			
Response (harmonised)	All	$D=0$	$D=1$	Diff. (*pp*)
Career (non-pecuniary)	14.79	14.52	14.95	0.43
No guarantee of a good job at the end	6.89	6.48	7.14	0.66
Don't need to go to university for the job someone may want	0.49	0.62	0.41	-0.20
Get less work experience	1.96	2.23	1.79	-0.44
Financial / debt	73.67	69.62	76.09	6.47
Now				
It is expensive	9.74	8.76	10.33	1.56
Not becoming financially independent	0.98	0.89	1.02	0.13
Not being able to start earning money / start work	6.34	6.54	6.21	-0.33
Costs (general / non specific)	23.18	21.14	24.38	3.24
Tuition fees / Accommodation costs / Living expenses	3.89	2.69	4.58	1.89
Future				
Getting into debt/have to borrow money	37.36	36.36	37.90	1.54
Social life / environment	9.37	9.75	9.14	-0.61
Leaving home / family / friends	8.16	8.90	7.72	-1.18
Stress	1.38	1.00	1.60	0.60
Education				
The workload can be hard / doubts about ability to finish course	6.00	5.37	6.38	1.00
Time	8.19	9.53	7.38	-2.15
Takes a long time	7.09	8.27	6.39	-1.88
Waste of time (general / non-specific)	1.18	1.34	1.09	-0.25

Notes: Only young people with at least 5 GCSEs at A*–C or equivalent were asked these questions ($N=3,469$). Where there are values in the right hand columns for the categories (in bold), these are the proportion of young people asked who mention at least one of the disadvantages in that category.

Figure 1.1: Proportion of students who mentioned specific advantages and disadvantages about going to university

(a) Harmonised *advantages*



Notes: Only students with at least 5 GCSEs at A*–C or equivalent were asked these questions ($N = 3,469$).

social lives. Interestingly, future graduates also seem to worry more about the potential financial downsides of attending university — perhaps suggesting that these are not a major barrier to university attendance in the UK.

As a first step towards comparing and quantifying the factors that potential students consider when deciding whether to attend university, we estimate a probability model to assess the predictive content in their reported beliefs. We started by estimating logit models with an indicator for holding a degree at age 25 the dependent variable, and all recorded advantages and disadvantages as (binary) indepedent variables. Estimates of key parameters are in appendix 2.E, table B2.[9] Many of the estimates are sizeable, but they are not very precisely estimated, as evidenced by the large standard errors. There are also a large number of estimates, which combined with their (im)precision, makes this model difficult to interpret.

To address these issues, we estimate a similar model using indicators for mentioning any response in each of the broader categories in tables 1.3 and 1.4 in place of the harmonised responses. The estimated parameters from this model are more straightforward to interpret. These are presented in table 1.5. The estimates in column (1) were obtained by regressing an indicator for holding a university degree at age 25 on indicators for mentioning any response from each of the broad categories of responses, and in column (2) we also include the following background characteristics as covariates: ethnicity, gender, A-levels, parental income, and a self-assessed ability measure.. The signs next to the categories in parentheses signal whether the responses in that category are described as advantages (+) or disadvantages (−).

The results of this exercise reflect our initial findings, with the categories with the biggest gaps between mentions by graduates and non-graduates in tables 1.3 and 1.4 having the largest coefficients. Importantly, even when we add a range of background controls in column (2) many of the coefficients remain sizable and statistically different from zero. This suggests that these questions are capturing variation in beliefs across individuals which impact their decision to attend university. The most important factors in the decision to attend university continue to be related to career (advantage), earnings (advantage), personal development (advantage), and the time it takes to get a degree (disadvantage).

The initial analysis presented in this section has highlighted the different factors young people consider when applying to university, and we have made a start at comparing their relative importance. However, although one of the responses we include in the model is about earnings, missing from our analysis so far is a proper measure of wages. Including earnings in our model will further benefit our analysis in (at least) two ways: (i) it will

[9] As these are qualitative survey responses, they are coded as indicator variables and are relative to a reference category, which is those who did not mention the corresponding advantage or disadvantage when surveyed. Also included are a range of background characteristics (ethnicity, gender, A-levels and parental income) for which we do not report the parameter estimates.

Table 1.5: Logit estimates (response categories)

Dependent variable:	Degree	
	(1)	(2)
Earnings	0.354***	0.294***
	(0.094)	(0.101)
Career (+)	0.658***	0.510***
	(0.081)	(0.088)
Career (−)	0.013	−0.054
	(0.102)	(0.110)
Financial (−)	0.169**	0.271***
	(0.084)	(0.094)
Social life (+)	0.272***	0.153
	(0.103)	(0.110)
Social life (−)	−0.250**	−0.121
	(0.127)	(0.138)
Education (+)	0.129	0.106
	(0.081)	(0.087)
Education (−)	0.197	0.354**
	(0.156)	(0.169)
Personal development (+)	0.617***	0.517***
	(0.102)	(0.108)
Time (−)	−0.281**	−0.253*
	(0.129)	(0.141)
Constant	−0.282***	−0.775***
	(0.102)	(0.223)
Observations	3,469	3,469
Log Likelihood	−2,237.408	−2,003.386
Akaike Inf. Crit.	4,496.816	4,088.772

Notes: *p<0.1; **p<0.05; ***p<0.01

Column (1) contains estimates from a logistic regression of an idndicator for holding a university degree at age 25 on indicators for mentioning each of the broad categories of responses in tables 1.3 and 1.4. Column (2) contains estimates for a regression also including the following background characteristics: ethnicity, gender, A-levels, parental income, and a self-assessed ability measure.

provide a better measure of the pecuniary benefits of attending university; (ii) comparing the contributions of other factors in the decision to wages will anchor these contributions to an interpretable metric.

1.3 Empirical framework

In this section we present a framework designed to allow us to compare and quantify the contributions of different factors in the decision to attend university, exploiting information on realised earnings, observed choices, and subjective beliefs. We build upon the analysis of the previous section by introducing a simple model which allows us to combine these different sources of data answer our research question: what is the relative importance of wages and non-pecuniary factors in the decision to attend university?

We start by introducing the key objects of the model and describing the behaviour of young people in our setup, along with some key assumptions we make to ensure our model is identified.

Utility of university or work. An individual's utility from choosing university ($s=1$), or work ($s=0$) is a linear combination of these different factors

$$U_{s,i} = \alpha Y_{s,i} + \theta'_{s,i}\gamma + Z'_{16,i}\delta_s + \epsilon_{s,i} \tag{1.1}$$

where Y_s represents the pecuniary factors (the logarithm of log weekly earnings in our application), θ_s is a vector of non-pecuniary (non-earnings) factors on which we have information, Z_{16} contains individual characteristics that may impact the non-pecuniary costs / benefits of university, and ϵ_s is a mean-zero random-utility term, all conditional on choice s.

Decision to attend university. At the time young people make their decision, they do not know the value that many of these outcomes will take, and so form expectations about their utility under each choice, based on their information set, $\mathcal{I}_i$:

$$\mathbb{E}[U_{s,i}|\mathcal{I}_i] = \mathbb{E}[\alpha Y_{s,i} + \theta'_{s,i}\gamma + Z'_{16,i}\delta_s + \epsilon_{s,i}|\mathcal{I}_i] \tag{1.2}$$

Individuals compare their expected utility of attending university, $U_1^{\mathcal{I}} (\equiv \mathbb{E}[U_1|\mathcal{I}_i])$, to that of working, $U_0^{\mathcal{I}}$, and choose the option with the higher expected utility. Therefore,

$$S \equiv \mathbb{1}\{U_1^{\mathcal{I}} - U_0^{\mathcal{I}} > 0\}. \tag{1.3}$$

This can be rewritten as the difference between expected (pecuniary) outcomes, and

expected "costs" of attending university, in the spirit of Roy (1951).

$$S \equiv \begin{cases} 1, & \text{if} \quad \alpha(Y_1^{\mathcal{I}} - Y_0^{\mathcal{I}}) + (\theta_1^{\mathcal{I}} - \theta_0^{\mathcal{I}})'\gamma + Z_{16}'(\delta_1 - \delta_0) + \epsilon_1^{\mathcal{I}} - \epsilon_0^{\mathcal{I}} > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{1.4}$$

This formulation leads naturally to an expression for the probability of attending university, conditional on expected earnings ($Y_s^{\mathcal{I}}$) and observed non-pecuniary factors (θ, Z_{16})

$$\Pr(S = 1|\mathcal{I}) = \Pr\left(\alpha(Y_1^{\mathcal{I}} - Y_0^{\mathcal{I}}) + (\theta_1^{\mathcal{I}} - \theta_0^{\mathcal{I}})'\gamma + Z_{16}'(\delta_1 - \delta_0) > \epsilon_0^{\mathcal{I}} - \epsilon_1^{\mathcal{I}}\right) \tag{1.5}$$

A chief aim of this paper is to estimate the relative importance of the pecuniary and non-pecuniary factors in the decision; i.e. how important is $\alpha(Y_1^{\mathcal{I}} - Y_0^{\mathcal{I}})$ versus $[(\theta_1^{\mathcal{I}} - \theta_0^{\mathcal{I}})'\gamma + Z_{16}'(\delta_1 - \delta_0)]$ when evaluating this conditional probability.

Expectations about (future) earnings. We need to specify exactly how young people form expectations about Y_s: what is in their information set, and how they use this information to form their expectations. We make the following assumptions:[10] (i) young people know the *true* process generating future incomes, but they only possess very limited information about the future — their information set reflects their current observable characteristics, i.e. X_{16}; (ii) young people only consider their earnings at age 25 (or these are a sufficient statistic for what they consider) when deciding whether to go to university.

Put differently, they are very good at predicting mean *realised* earnings among their peers conditional on X_{16}, but they are not very good at predicting their own *future* characteristics (or they do not know how earnings depend on their future characteristics). Then $Y_s^{\mathcal{I}} \equiv \mathbb{E}[Y_s \mid X_{16}]$. Under these assumptions earnings expectations, $Y_s^{\mathcal{I}}$ are identified from *realised* earnings, and the students' characteristics at 16. As we only observe either Y_1 or Y_0 for each individual, these assumptions allow us to estimate expected wages $Y_1^{\mathcal{I}}$ and $Y_0^{\mathcal{I}}$ for young person, and hence an expected graduate wage premium, $Y_1^{\mathcal{I}} - Y_0^{\mathcal{I}}$.

Expectations about other (non-pecuniary) factors. We use the harmonised responses to open-ended questions about the advantages and disadvantages of going to university, discussed in detail in section 1.2, to measure the expected other factors premium, $\theta_1^{\mathcal{I}} - \theta_0^{\mathcal{I}}$. Limited somewhat by the nature of these questions, we assume that individuals

[10]The current gold standard are elicited expectations about earnings, following the advice of Manski (1993). However, these are rare, especially in large representative samples. Our assumption that young people know the true income process is standard in economics. For example, Cunha and Heckman (2007) assume this, and develop a method for testing the contents of young people's information sets. Willis and Rosen (1979) and D'Haultfoeuille and Maurel (2013) also assume a similar model for expected earnings, though they allow for an unobserved component in the young people's information set, i.e. $\mathbb{E}[Y_s|X_{16}, \eta_1, \eta_0]$, where η_s are not observed by the econometrician. We assume that we observe all the information that young people use to form their expectations about earnings. We plan to compare our model (and subsequent results) with other models of expectations / information sets in future work.

either believe there to be no difference in this factor whether they go to university or not, or they believe there will be a difference, which is fixed to be of constant size across all individuals who hold this belief. Therefore for each factor mentioned by *any* student, the component of $\theta^{\mathcal{I}}_{1,i} - \theta^{\mathcal{I}}_{0,i}$ takes one value (normalised to 1) if mentioned by student i, and another value (normalised to 0) if not mentioned. The parameter γ_j on the j-th component of $\theta^{\mathcal{I}}_{1,i} - \theta^{\mathcal{I}}_{0,i}$ then reflects (average) preferences for this aspect of university. We also allow the non-pecuniary factors to vary with individual characteristics, captured by the vector Z_{16}.

1.3.1 Identifying the parameters in the utility function

Recall the probability of attending university, conditional on expectations about earnings and other factors, in the model:

$$\Pr\left(S=1|Y^{\mathcal{I}}_1 - Y^{\mathcal{I}}_0, \theta^{\mathcal{I}}_1 - \theta^{\mathcal{I}}_0, Z_{16}\right) = \Pr\left(\alpha(Y^{\mathcal{I}}_1 - Y^{\mathcal{I}}_0) + (\theta^{\mathcal{I}}_1 - \theta^{\mathcal{I}}_0)'\gamma + Z'_{16}\delta > \epsilon^{\mathcal{I}}_0 - \epsilon^{\mathcal{I}}_1\right). \tag{1.6}$$

Identification of α, $\delta \equiv \delta_1 - \delta_0$ and γ then requires assumptions on the distribution of the random-utility terms, ϵ_1 and ϵ_0. A standard assumption in the discrete-choice literature is that these follow a type-I extreme-value distribution, meaning their difference follows a logistic distribution: $(\epsilon_0 - \epsilon_1) \sim$ Logit. The parameters α, δ and γ capture the relative contribution of earnings and observed other factors to young people's utility, and hence in their decision to attend university. These parameters are only identified up to a scale normalisation.

1.3.2 Estimation

Having laid out the assumptions we make to identify our model, we now describe our estimation strategy.

Expected graduate-wage premium, $Y^{\mathcal{I}}_1 - Y^{\mathcal{I}}_0$. Under our model for young people's expectations, the expected earnings we need are $\mathbb{E}[Y_s|X_{16}]$. Given X_{16} we use OLS to estimate this conditional expectation.[11] We estimate the simplest linear conditional expectation for each level of education, with no interactions. We then use the estimated

[11] By using OLS we do not control for selection. Young people base their decision on their expected graduate and non-graduate earnings, $Y^{\mathcal{I}}_s$. Our assumption about how they form these expectations means that we have access to the same information they do—therefore, there is not selection on unobservables. As mentioned in a previous footnote, we plan to test different models of assumptions in future work. We did attempt to estimate different wage equations allowing for selection, which we present in the appendix, figure C1. When assuming normal errors following Heckman (1979) we found that for the majority of CMs their graduate earnings premium $\widehat{Y}^{\mathcal{I}}_1 - \widehat{Y}^{\mathcal{I}}_0$ was negative, suggesting that the normality assumption is problematic. Estimating a two-stage model with a more flexible first-stage (using Coppejans (2001) mixture-of-distributions (MOD) estimator) resulted in an almost identical observed graduate wage premium distribution.

coefficients, $\hat{\beta}_{s,16}$, to obtain estimates $\widehat{Y}_s^{\mathcal{I}} = X_{16}'\hat{\beta}_{s,16}$. The estimated expected graduate-wage premium is simply $\widehat{Y}_1^{\mathcal{I}} - \widehat{Y}_0^{\mathcal{I}} = X_{16}'(\hat{\beta}_{1,16} - \hat{\beta}_{0,16})$. We include the following covariates in X_{16}: parents' occupations, parents' education level, a measure of parental income, the number of A-levels a student is taking, gender, and whether high pay is important to them.

The parameters of the utility function, α, δ and γ. We estimate the parameters of the utility function using logistic regression. To avoid perfect multicollinearity when estimating equation (1.6), X_{16} must not be a subset of Z_{16}. An alternative would be to transform log-wages by some (non-linear) function. We choose to exclude beliefs relating to earnings from Z_{16}.

Distributions of earnings and other factors. The aim of this paper is to compare and to quantify the roles of earnings and non-pecuniary factors in the decision to attend university. To do this, we estimate comparable distributions of the different factors using the following strategy: (i) obtain estimates $\hat{\alpha}$, $\hat{\gamma}$, and $\widehat{Y}_1^{\mathcal{I}} - \widehat{Y}_0^{\mathcal{I}}$; (ii) recombine these estimates with the data $(X_{16}, \theta_1^{\mathcal{I}} - \theta_0^{\mathcal{I}}, Z_{16})$, to calculate the "contribution" of each (type of) factor; (iii) transform these utility values to be equivalent to a difference in log-earnings; (iv) use a kernel-density estimator to plot the empirical distributions.

1.4 Results

In this section we present and discuss the results of estimating the model described in section 3.2. We first present results for the full sample, and then study variation across different socio-economic groups and over time.

1.4.1 Main results

Kernel density estimates of the distributions of earnings premiums (blue solid line) and other factors premiums (red dashed line) are presented in figure 1.2. Table 1.6 presents further statistics on these distributions: their mean, first and last deciles, and quartiles. The locations of the two distributions are remarkably similar, evidenced by their similar means: 8 log-points (earnings) and 6 log-points (non-earnings). However, the dispersion of the other factors premium is much higher. This difference is visible both in figure 1.2, and by comparing the interquartile ranges in table 1.6. The interquartile range is 14 log-points for the earnings premium, while it is over twice as large at 30 log-points for the non-earnings factors. The same is true of the interdecile range (26 versus 59 log-points), and the standard deviation (10 versus 24 log-points).

Table 1.6: Summary statistics for the earnings and other factors premium distributions

	Mean	$q_{.1}$	$q_{.25}$	$q_{.5}$	$q_{.75}$	$q_{.9}$
Total	0.14	-0.15	-0.01	0.14	0.30	0.43
Earnings	0.08	-0.05	0.01	0.08	0.15	0.21
Non-earnings	0.06	-0.23	-0.09	0.06	0.21	0.36
Career	0.03	0.00	0.00	0.05	0.05	0.05
Financial	0.05	0.00	0.00	0.06	0.06	0.07
Education	0.01	-0.01	0.00	0.00	0.00	0.06
Personal dev.	0.02	0.00	0.00	0.00	0.00	0.06
Social life	0.00	0.00	0.00	0.00	0.00	0.02
Other	0.18	-0.09	0.04	0.17	0.31	0.45

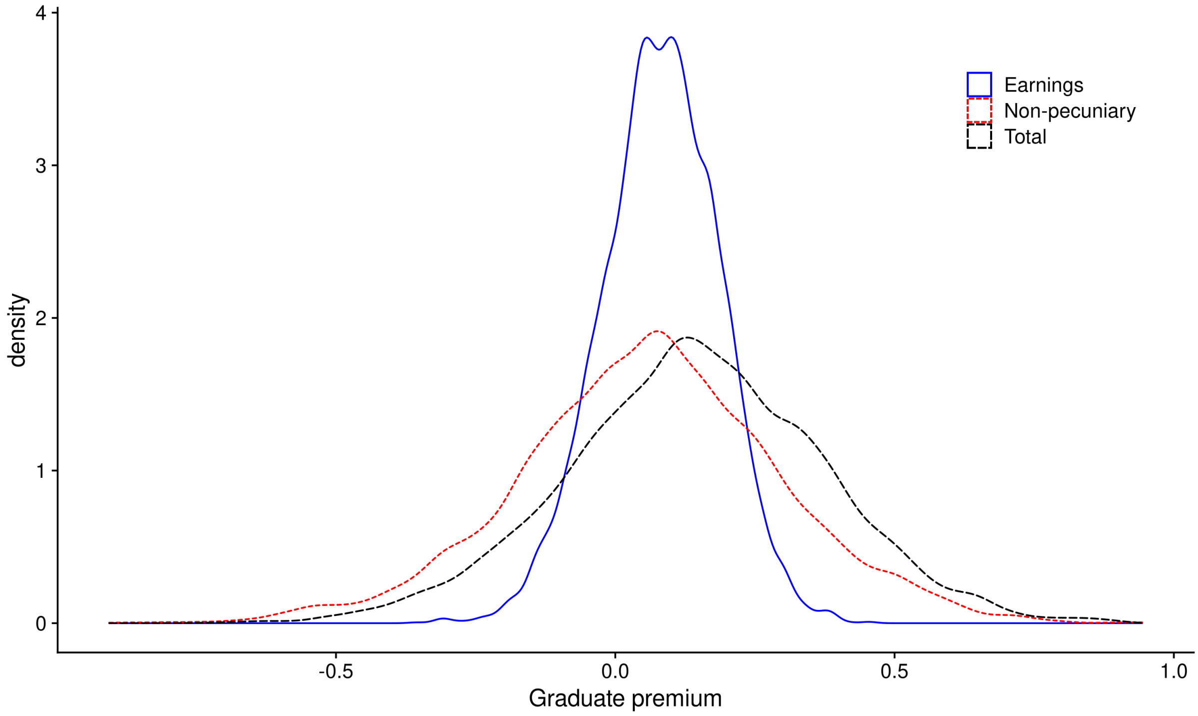
Notes: The values for "Time" are omitted as they do not vary between the 1st and 9th deciles. The values are in units equivalent to a difference in log-wages.

We can also compare the number of young people who have positive values of each factor. Among graduates, 80% have a positive value of pecuniary factors and 73% of non-pecuniary factors. Therefore, 20% of graduates still attend university despite expecting negative pecuniary returns. For non-graduates, 76% expect positive pecuniary returns, though only 42% expect positive non-pecuniary returns. Therefore, it appears to be chiefly the influence of these non-earnings factors that determines whether a young person decides to attend university, a role reflected in the similarity between the distribution of all factors in the decision (black long-dashed line, figure 1.2) and the other factors (red dashed line, figure 1.2).

Counterfactual exercise

In order to further highlight the importance of non-earnings factors, we perform the following counterfactual exercise, borrowed from D'Haultfoeuille and Maurel (2013). We calculate the predicted probabilities of university attendance under different (fixed across the sample) values of the pecuniary and non-pecuniary factors. The results of this exercise are in table 1.7. If everyone in the sample had other factors equal to the 10th percentile value, only 35% of people would attend university—over 27*pp* fewer than did actually attend. Moreover, assigning everyone other factors equal to the 90th percentile results in over 86% of people attending university, an increase of over 20*pp*. Conversely, varying the expected graduate-wage premium between the 10th and 90th percentiles has a much smaller effect on university attendance. Over 50% of young people would still attend if they expected a pecuniary premium equal to the 10th percentile, while 73% would attend if we fix all expectations about the pecuniary gains at the 90th percentile. This emphasises the key role non-pecuniary factors play in the decision to attend university.

Figure 1.2: Distributions of earnings and other factors premiums in the decision to go to university



Notes: The values of the factors are estimated as described in 2.3. The distributions are then estimated (and plotted) with the kernel density estimator in the `R` package `ggplot2`, using the default Gaussian kernel and bandwidth (Wickham, 2016).

Table 1.7: University attendance under counterfactual factor values

Counterfactual	Earnings	Other	University (%)
Data	0.08	0.06	62.2
Earnings			
10th percentile	-0.053	-	50.8
25th percentile	0.014	-	56.8
75th percentile	0.152	-	68.4
90th percentile	0.210	-	72.7
Other			
10th percentile	-	-0.233	34.7
25th percentile	-	-0.087	49.4
75th percentile	-	0.212	77.2
90th percentile	-	0.357	86.1

Notes: The "units" of earnings and other factors are equivalent to a difference in log-wages. University is the fraction who attend under the counterfactual distribution. The values in row "Data" are: the median values of earnings and other factors premiums.

Decomposing non-pecuniary factors.

So far our analysis has considered all non-earnings factors together. However, using information on young people's beliefs, we can attempt to decompose these non-pecuniary factors. We do so by calculating the portion of non-pecuniary returns attributable to variation in a given aspect of university. For example, the following aspects of university mentioned by students are related to their career: get better job, better opportunities, need for career, show skills, delay get job, not earning / working, no job guarantee, not needed for job, and less experience. Therefore, we can calculate the expected values of non-pecuniary career-related returns using these variables and their estimated coefficients.

The summary statistics for the distributions of non-pecuniary returns associated to career, financial, educational, personal development, social life, and other are in table 1.6. The allocation of responses to each category is detailed in table B1, and also corresponds to the categorisation in tables 1.3 and 1.4. The category "time" is omitted from table 1.6 as the values of this factor did not vary between the 10th and 90th percentiles. The variation in "other" is due to variation in non-pecuniary factors captured by the included background characteristics. Although some varation in non-pecuniary returns is associated with the recorded beliefs, the majority is due to these background characteristics. Given the limited variation in the indicator variables we rely upon to measure beliefs, it is likely that the variation attributed to each category in table 1.6 represents a lower bound for the true contributions of these aspects to the non-pecuniary returns to university.

Nevertheless, we can still say something about the contributions to non-pecuniary returns for some of these aspects of (life at and after) university. For both career-related and financial non-pecuniary returns, the 25th percentile value is zero, while the median is 5 log points. This variation is similar to that of pecuniary returns, which range from 1 log-point at the 25th percentile to 8 log-points at the median. Therefore, despite failing to explain the majority of variation in non-pecuniary returns, the portion of these returns which is explained by variation in beliefs is still sizable compared to expected pecuniary returns.

Limitations

The analysis in this paper has a number of limitations, which we will begin to discuss here. First, we rely on strong assumptions about how young people form expectations about the pecuniary returns to university, both in terms of what measure of realised earnings they base they use to form these expectations, and on what information they include in their information set when forming these expectations. It is unclear exactly how these assumptions are likely to have impacted our results.

For example, if young people are less myopic than we assume, and consider their lifetime earnings, then their expectations will also depend on the relative growth rates

of graduate vs non-graduate wages. Then, if graduate wage growth is higher than non-graduate, our results would *underestimate* the expected pecuniary returns to university. However, we do not have full realised lifetime earnings for this cohort as they are still at the start of their careers, so would have to make additional assumptions on how they form expectations about this growth rate. Regarding the contents of young people's information sets, it is possible we do not observe all the information that young people use to form their expectations about the pecuniary and non-pecuniary returns. Recent work has developed methods to allow unobserved components in the expected pecuniary returns (D'Haultfoeuille and Maurel, 2013), and to determine the contents of young people's information sets Cunha et al. (2004). We plan to test different models of expectations in future work, including allowing for unobserved heterogeneity in both pecuniary and non-pecuniary returns.

Finally, although we have started to decompose the non-pecuniary returns into meaningful components, the majority of these returns are still unattributed. Therefore, in future work it will be important to use (where available), and collect, more detailed data on young people's expectations and beliefs about the non-pecuniary aspects of university.

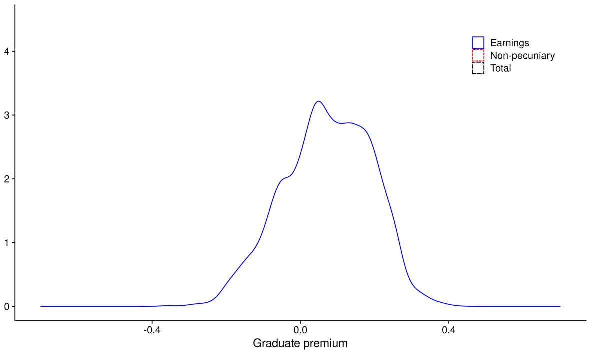
1.4.2 Results by socio-economic status (SES)

In this section we present the distributions of earnings and other factors premiums *conditional on socio-economic status*. We use parental earnings at age sixteen as a measure of socio-economic status (SES). Comparing the factor distributions across SES allows us to quantify the relative contributions of earnings and other factors to the SES-gap in university attendance (see table 1.2). The SES-gap in education in the UK is a name for the finding, documented by many researchers, that young people from less advantaged backgrounds are much less likely to attend university. Even in our sample, which includes only higher ability students, the SES gap in university attendance is 15 percentage points (*pp*), with 60% of those in the low and middle SES groups (bottom and middle three quintiles of parental income) attending university, compared with 75% in the top SES group (top quintile).

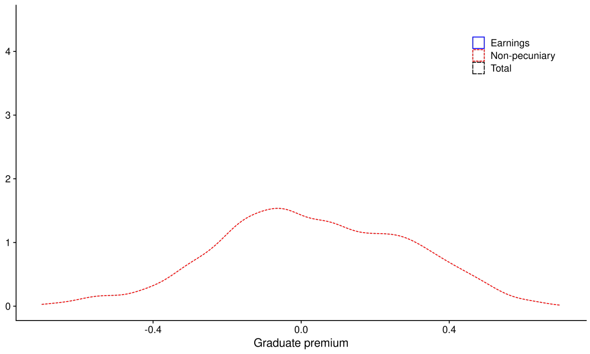
Figure 1.3 shows the distributions of earnings (left column) and other (right column) factors, for those with parents in the bottom 20% (top row), middle 60% (middle row), and top 20% (bottom row) of the earnings distribution. Focusing first on earnings (left column, figure 1.3), the distributions of factors across the three groups are similarly located, though the means are slightly increasing in parental income (table 1.3g). For other factors (right column, figure 1.3), the distributions across the three groups clearly occupy different locations, and their means are strongly increasing in parental income. The mean other factors in the bottom and middle SES groups are slightly positive while they are strongly positive for the top SES group. The SES-gap in educational attainment

Figure 1.3: Comparing factor distributions by parental income (SES)

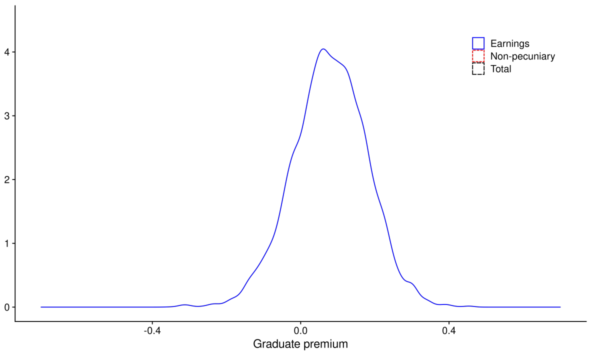
(a) Earnings, bottom 20%



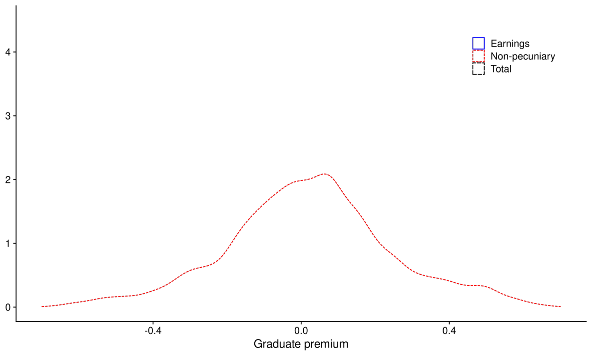
(b) Other, bottom 20%



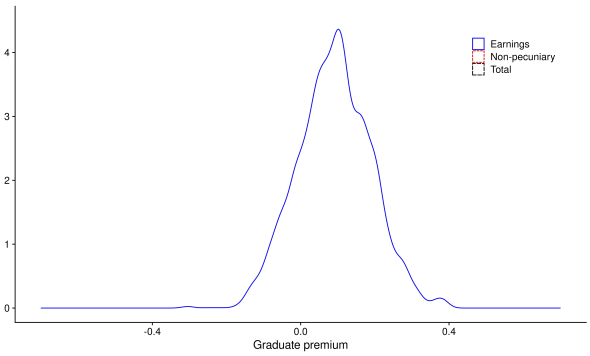
(c) Earnings, middle 60%



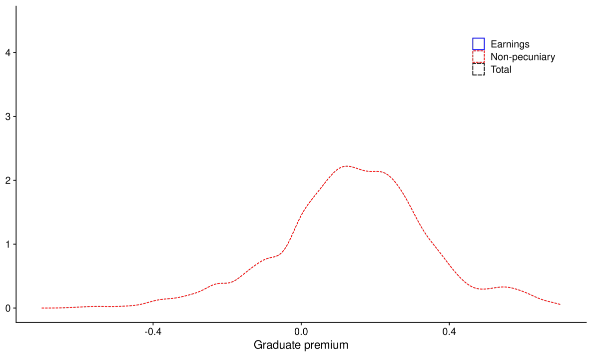
(d) Other, middle 60%



(e) Earnings, top 20%



(f) Other, top 20%



(g) Summary statistics for the distributions in panels (a)–(f)

	Mean	$q_{.1}$	$q_{.25}$	$q_{.5}$	$q_{.75}$	$q_{.9}$
Earnings						
Bottom 20%	0.07	-0.09	-0.01	0.07	0.16	0.22
Middle 60%	0.08	-0.05	0.01	0.08	0.15	0.20
Top 20%	0.09	-0.04	0.03	0.10	0.16	0.21
Other						
Bottom 20%	0.03	-0.29	-0.15	0.02	0.23	0.37
Middle 60%	0.02	-0.26	-0.11	0.02	0.15	0.32
Top 20%	0.15	-0.11	0.03	0.15	0.27	0.39

Notes: The values of the factors are estimated as described in 2.3. The distributions are then estimated (and plotted) with the kernel density estimator in the `R` package `ggplot2`, using the default Gaussian kernel and bandwidth (Wickham, 2016). The mean and standard deviations in panel (g) are in %Δ wage equivalent.

is mostly driven by other factors in our analysis.

These findings are broadly inline with recent work by Boneva and Rauh (2020) who find that a large part of the gap between high and low SES students is due to differences in other factors premium. They also find a similarly sized role for wage premium, for which we find a much smaller role. There are a number of differences between our analyses that could explain this discrepancy. First, as we include only those individuals asked specifically about their beliefs in our sample, we are forced to focus on higher ability students. Second, Boneva and Rauh (2020) directly elicit expectations about earnings, while we estimate these expectations from realised earnings. Third, our definitions of SES are quite different, as Boneva and Rauh (2020) do not have detailed information on the parents of their sample members, and so define SES based on parental education, while we use parental income. Nonetheless, our findings add to the growing evidence that differences in (beliefs about) the non-pecuniary aspects of university across different groups are key drivers of differences in educational attainment across these groups.

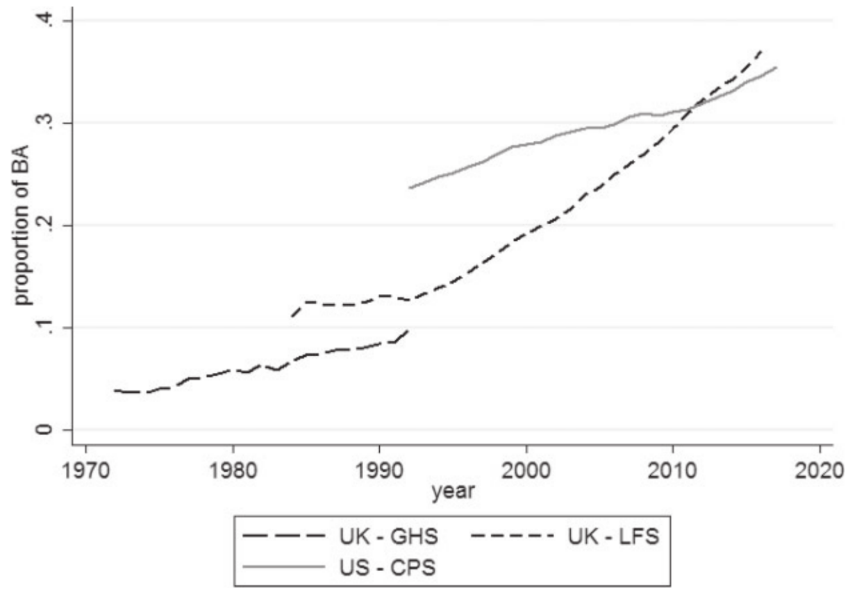
1.4.3 Changes over two decades

In an effort to shed light on what drove more and more people to attend university in England in recent decades, despite apparent stagnation in the wage returns, we re-estimate our model on data from a cohort born in 1970. These trends are displayed in figure 1.4, reproduced from Blundell et al. (2022). In figure 1.4a the trend in proportion of those aged 30 with a first degree (BA) is plotted for cohorts born between 1950 and 1985, for the UK (blue line) and US. The level of educational attainment grew much faster for the UK over this period. In figure 1.4b the ratio of median BA to high school wages is plotted for the same period, which is remarkably flat. Taken together, these facts suggest that there must have been a large increase in (expected) non-pecuniary returns to explain the growth in higher education over this period. We test this hypothesis, comparing cohorts born in 1970 and 1990.

The data for the earlier cohort is from the British Cohort Study 1970 (BCS 1970), a similar study to *Next Steps* which follows all 16,000 people born in the UK in a single week in April 1970. The aims of the BCS 1970 are very similar to *Next Steps*, and therefore we have very similar information on the cohort members. In particular, they were interviewed at age 16, when they were asked a series of questions about their expectations for the future, and we also have information on their family background at this point. They were then interviewed again 10 years later at age 26, with their earnings and their qualifications the key information we use from that wave. Having very similar data from two cohorts born 20 years apart allows us to directly compare the distributions of the earnings and other factors premiums we estimate for these two cohorts. We estimate the model on the earlier cohort following the procedure described in section 2.3.

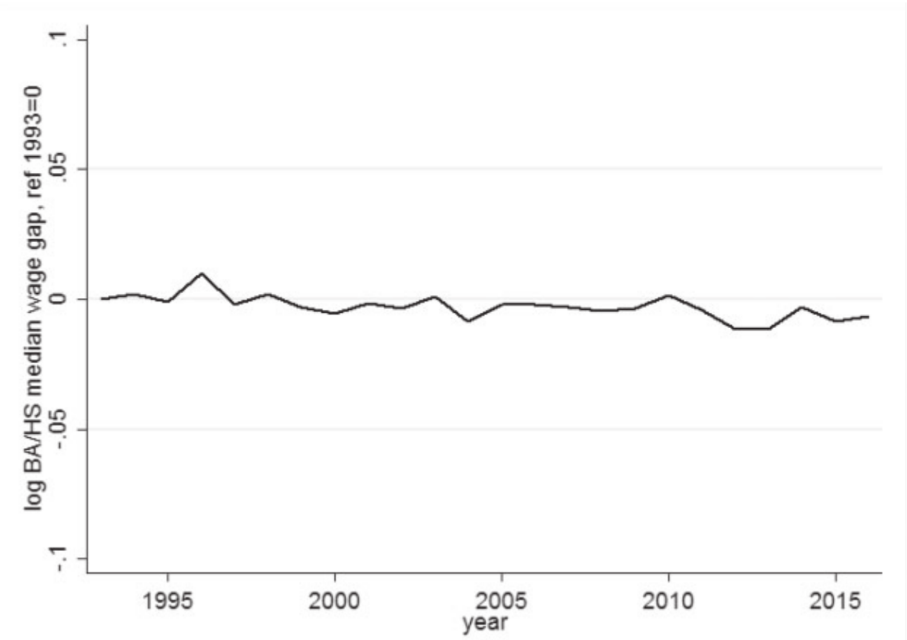
Figure 1.4: Higher education and wages in the UK vs the US in recent decades

(a) Proportion of people with a BA or higher education by cohort, UK and US



Source: Reproduced from Blundell et al. (2022). Those authors' calculation from the U.K. Labour Force Survey, the U.K. General Household Survey, and the U.S. Current Population Survey.
Notes: BA refers to individuals who have a bachelors or higher degree. Blundell et al. (2022) aggregate each dataset to the level of year and 5-year age band, and regress the BA proportion on year dummies and age-band dummies. The proportion BA numbers are year effects from these regressions plus the level in 1992 for the 30–34 age band.

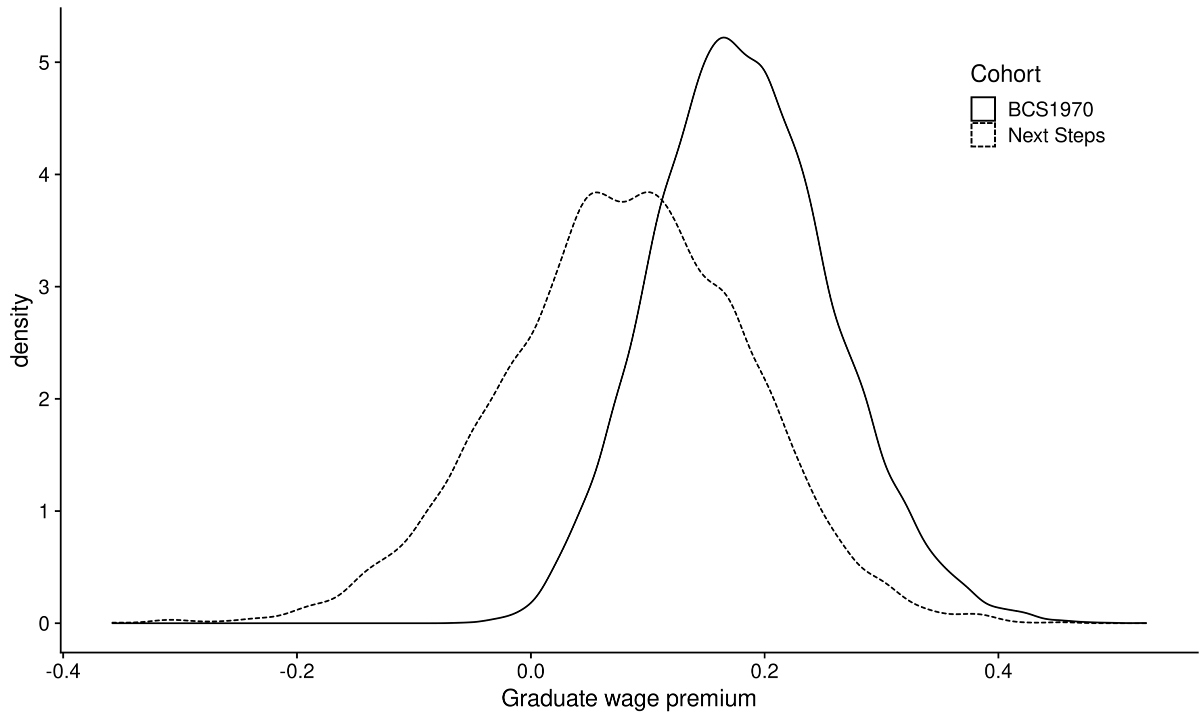
(b) Ratio of BA median wage to that of high-school graduates 1993–2016, U.K.



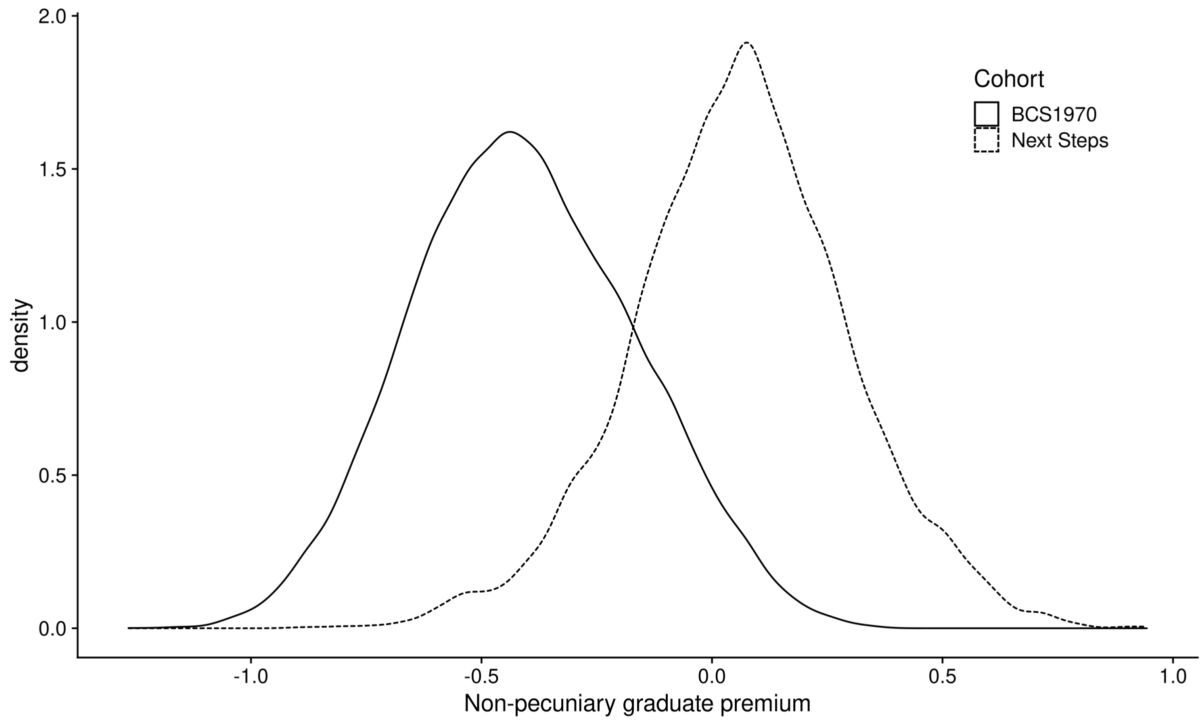
Source: Reproduced from Blundell et al. (2022)
Notes: Wage is hourly. The sample is 20–59 year olds in LFS 1993–2016. BA refers to individuals who have a bachelors or higher degree. Blundell et al. (2022) aggregate LFS to the level of year and 5-year age groups, and regress the log BA to HS median wage ratio on year dummies and age-band dummies. The figure plots the estimated year effects normalized to zero in 1993.

Figure 1.5: Changes in distributions of factors between cohorts (1970–1990)

(a) Earnings



(b) Other



Source: The data are from the BCS1970 and Next Steps surveys.

Notes: The values of the factors are estimated as described in 2.3. Only young people with at least 5 GCSEs at A*–C or equivalent are included in the sample. The distributions are then estimated (and plotted) with the kernel density estimator in the `R` package `ggplot2`, using the default Gaussian kernel and bandwidth (Wickham, 2016).

Table 1.8: Comparison of premiums between cohorts

Cohort:	BCS (1970)	LSYPE (1990)	Change 1970–1990
Degree	29.9%	62.6%	32.7*pp*
Earnings			
10th percentile	0.082	-0.053	-0.140
25th percentile	0.124	0.014	-0.115
50th percentile	0.175	0.083	-0.096
75th percentile	0.231	0.152	-0.080
90th percentile	0.281	0.210	-0.073
Other			
10th percentile	-0.639	-0.233	0.483
25th percentile	-0.500	-0.087	0.492
50th percentile	-0.334	0.065	0.481
75th percentile	-0.150	0.212	0.451
90th percentile	0.003	0.357	0.440

Notes: The percentiles are expressed in units equivalent to the difference in log wages.

Figure 1.5 presents the estimated distributions of earnings and other factors premiums for the two cohorts. The mean graduate-wage premium *decreased* on average between those born in 1970 (solid blue line) and 1990 (dashed blue line). Meanwhile, the other factors premium *increased* significantly on average over this period. Key percentiles of these distributions and their differences are in table 1.8, and we can see that the median earnings premium fell by 9.6 log-points, while the median non-pecuniary factors premium increased by 48.1 log-points. Recall, the units of these premiums are equivalent to a difference in log-wages. In the 1970 cohort, the median cohort member believed their wages would be over 16% higher if they attended university, while the median 1990 cohort member believed university would increase their wages by 9.2%. However, the median 1970 cohort member perceived significant non-earnings "costs" to attending university, equivalent to 28.8% of their earnings. These costs had become negative (i.e. benefits) for the later cohort, who perceived non-earnings benefits of attending university equivalent to 6.3% of their earnings.

Therefore, in our framework the large increase in higher education attainment between the two cohorts (see table 1.8) was entirely driven by an increase in expectations about non-earnings factors. This finding is inline with the evidence, presented at the beginning of this section, that the pecuniary returns to a degree have remained remarkably constant over a period of significant growth in higher education in the UK.

1.5 Conclusion

In this paper we specify and estimate a model of educational choice, that specifically includes expectations about earnings and other, financial and non-pecuniary, factors. We exploit data on a cohort born at the end of the 1980s which features data on realised earnings and expectations about the non-pecuniary costs and benefits of going to university. Our findings add support to the notion that individuals are not strict income maximisers when they make educational choices. We find that the non-pecuniary premium is able to explain most of the variation across individuals that causes some people to attend university and others to not, with the graduate-earnings premium playing only a minor role. Splitting the sample by parental income (a measure of socio-economic status), we find that differences in factors other than earnings across socio-economic groups are chiefly responsible for the "SES gap" in educational attainment. Finally, comparing the roles of pecuniary and other factors in educational decisions across a period of significant growth in higher education attainment and increased financial costs, we find that the expected graduate premium fell slightly, suggesting increases in the value of non-pecuniary factors drove the expansion in attainment.

Although further work is required to address the limitations of our analysis, our results suggest that a better understanding of the (expected) non-pecuniary costs and benefits of university is vital. Of particular importance is the apparent difference between the most-advantaged young people and their less-advantaged peers in terms of their expectations about the non-pecuniary returns to university. The socio-economic status gap in educational attainment is a barrier to social mobility and contributes to inequality: both in terms of income, and other outcomes known to be related to education such as health.

In future work, we plan to extend the analysis in this paper in a number of directions. First, by using different models of expectations we can test the sensitivity of our finding to our assumptions about how young people form their expectations. Comparing these estimated expectations under different models with the gold-standard of elicited expectations will be an important part of this work. Second, although we have started to decompose the so-called "psychic costs" of university into meaningful components, the limitations of our data mean there is still lots of work to do in this area. This will require both careful analyis of existing data, including from more recent cohort studies in the UK, as well as original data collection.

Bibliography

Anders, J. (2012): "The Link between Household Income, University Applications and University Attendance*," *Fiscal Studies*, 33, 185–210.

Anders, J. and J. Micklewright (2015): "Teenagers' Expectations of Applying

to University: How do they Change?" *Education Sciences*, 5, 281–305, number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

ARCIDIACONO, P., V. J. HOTZ, AND S. KANG (2012): "Modeling college major choices using elicited measures of expectations and counterfactuals," *Journal of Econometrics*, 166, 3–16.

ARCIDIACONO, P., V. J. HOTZ, A. MAUREL, AND T. ROMANO (2020): "Ex Ante Returns and Occupational Choice," *Journal of Political Economy*, 128, 4475–4522.

BLANDEN, J. AND S. MACHIN (2004): "Educational Inequality and the Expansion of UK Higher Education," *Scottish Journal of Political Economy*, 51, 230–249.

BLUNDELL, R., D. A. GREEN, AND W. JIN (2021): "The UK as a Technological Follower: Higher Education Expansion, Technological Adoption, and the Labour Market," *Review of Economic Studies*, 81.

——— (2022): "The U.K. as a Technological Follower: Higher Education Expansion and the College Wage Premium," *The Review of Economic Studies*, 89, 142–180.

BONEVA, T. AND C. RAUH (2020): "Socio-Economic Gaps in University Enrollment: The Role of Perceived Pecuniary and Non-Pecuniary Returns," *HCEO Working Paper*, 83.

BRITTON, J., L. VAN DER ERVE, C. BELFIELD, L. DEARDEN, A. VIGNOLES, M. DICKSON, Y. ZHU, I. WALKER, L. SIBIETA, AND F. BUSCHA (2021): "How much does degree choice matter?" Tech. rep., The IFS.

COPPEJANS, M. (2001): "Estimation of the binary response model using a mixture of distributions estimator (MOD)," *Journal of Econometrics*, 102, 231–269.

CRAWFORD, C. AND W. M. JIN (2014): "Payback time? Student debt and loan repayments: what will the 2012 reforms mean for graduates?" Tech. rep., Institute for Fiscal Studies.

CUNHA, F. AND J. J. HECKMAN (2007): "Identifying and Estimating the Distributions of Ex Post and Ex Ante Returns to Schooling," *Labour Economics*, 14, 870–893.

CUNHA, F., J. J. HECKMAN, AND S. NAVARRO (2004): "Separating uncertainty from heterogeneity in life cycle earnings," *Oxford Economic Papers*, 57, 191–261.

DEARDEN, L., E. FITZSIMONS, AND A. GOODMAN (2005): "Higher education funding policy: a guide to the debate," .

Delavande, A. and B. Zafar (2019): "University Choice: The Role of Expected Earnings, Nonpecuniary Outcomes, and Financial Constraints," *Journal of Political Economy*, 127, 2343–2393, publisher: University of Chicago.

D'Haultfoeuille, X. and A. Maurel (2013): "Inference on an extended Roy model, with an application to schooling decisions in France," *Journal of Econometrics*, 174, 95–106.

Dominitz, J. and C. F. Manski (1996): "Eliciting Student Expectations of the Returns to Schooling," *The Journal of Human Resources*, 31, 1–26.

Gong, Y., T. Stinebrickner, and R. Stinebrickner (2019): "Uncertainty about future income: Initial beliefs and resolution during college," *Quantitative Economics*, 10, 607–641, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/QE954.

Heckman, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153.

Heckman, J. J., J. E. Humphries, and G. Veramendi (2018): "Returns to Education: The Causal Effects of Education on Earnings, Health, and Smoking," *Journal of Political Economy*, 126, S197–S246, publisher: The University of Chicago Press.

Heckman, J. J., L. J. Lochner, and P. E. Todd (2006): "Chapter 7 Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond," in *Handbook of the Economics of Education*, Elsevier, vol. 1, 307–458.

IOE Centre for Longitudinal Studies (2008): "Next Steps: Wave 4 Documentation," .

——— (2018): "Next Steps: Sweeps 1–8. 2004–2016. [data collection]." .

Manski, C. F. (1993): "Adolescent Econometricians: How Do Youth Infer the Returns to Schooling?" *NBER*, 43–60.

Mcfadden, D. (2001): "Economic Choices," *THE AMERICAN ECONOMIC REVIEW*, 91, 53.

OECD (2018): *Equity in Education: Breaking Down Barriers to Social Mobility*, PISA, OECD.

Oreopoulos, P. and U. Petronijevic (2013): "Making College Worth It: A Review of the Returns to Higher Education," *The Future of Children*, 23, 41–65.

Oreopoulos, P. and K. G. Salvanes (2011): "Priceless: The Nonpecuniary Benefits of Schooling," *Journal of Economic Perspectives*, 25, 159–184.

ROY, A. D. (1951): "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 3, 135–146.

WALKER, I. AND Y. ZHU (2008): "The College Wage Premium and the Expansion of Higher Education in the UK," *The Scandinavian Journal of Economics*, 110, 695–709.

WICKHAM, H. (2016): *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.

WILLIS, R. J. AND S. ROSEN (1979): "Education and Self-Selection," *Journal of Political Economy*, 87, S7–S36.

WISWALL, M. AND B. ZAFAR (2015): "Determinants of College Major Choice: Identification using an Information Experiment," *The Review of Economic Studies*, 82, 791–824.

WOODIN, T., G. MCCULLOCH, AND S. COWAN (2013): *Secondary Education and the Raising of the School-Leaving Age: Coming of Age?*, Secondary Education in a Changing World, Palgrave Macmillan US.

1.A Institutional context of HE in England

In this section we discuss the organisation of higher education in England. Schooling is compulsory up to the age of sixteen in the UK, and has been since 1972 (Woodin et al., 2013). Figure E1 presents the time-line of decisions and exams that students (generally) must take to secure a place at university. Two key decisions are: the application to continue on to further education ("sixth form") in the final year of secondary school; and the university application in the final year of sixth form. The main data source follows individuals through secondary school and beyond, from the age of 14 until 19. However, in this paper we will focus exclusively on the decision to attend university and treat the outcome of the decision to continue to sixth form as a predetermined characteristic. Estimating a dynamic discrete-choice model to exploit more of the data is an interesting avenue we hope to explore in future work.

University application process. The UK university application system is quite unique in many ways, and is worthy of study in its own right. Students apply through a centralised system, the "Universities and Colleges Admissions Service" (UCAS)[12] in the autumn of their final year of sixth form. Students can apply for up to five places, where each "place" is a *university-subject pair*. The application consists of a personal statement written by the student, predicted A-levels grades from their teachers, and past national-exam results. These are common across all applications, so students cannot tailor their personal statement to different subjects or institutions.[13] Students then receive *conditional* offers or are rejected from each place they applied, and must select two of their offers: a first choice and a back-up option. The offers made to students in sixth form are (almost exclusively) conditional on their future grades, so for example may require a student sitting 3 A-levels to achieve AAB, with one A in chemistry. The back-up option allows the student to aim high with their first choice, and still have a place somewhere if they fail to achieve those grades. Students sit their A-levels knowing their required grades for each place, and are automatically accepted at their first choice if they achieve the required grade, at their second if they miss the requirement for their first choice, and nowhere if they do not meet either requirement.[14]

The funding of higher education. Universities in the UK are privately run, but receive state funding *and* are regulated by government over the fees they can charge their students. Tuition fees were first introduced for UK students at UK universities in 1998. Prior to this, universities could not charge fees for tuition. There was also a

[12]Universities and colleges are different entities in the UK, and the names are not used interchangeably, unlike in the US.

[13]This is an implicit barrier which stops people applying to vastly different subjects.

[14]There is a mechanism to allocate students who missed their offers on both their first- and second-choices to places at universities who remain unfilled called "Clearing".

Figure A1: Timeline of educational decisions (1990 cohort)

Year 11 (age 15/16)	Apply for sixth form
End of year 11	Sit GCSEs & legally able to leave school
End of year 12	Sit AS-levels
Year 13 (age 17/18)	Apply to universities & receive "conditional" offers
End of year 13	Sit A-levels
Summer after year 13	Receive A-levels results & confirm place at university

system of grants and loans in place to cover living costs. In 1998 a means-tested fee was introduced, with the students from the most privileged backgrounds paying £1,000 per year in tuition fees. The poorest students were entitled to a 25% reduction. The situation changed again in 2006, with the introduction of so-called "top-up" fees, which could be set by each university up to a maximum of £3,000.[15] Alongside these fees, the government introduced a comprehensive system of loans and grants to cover both tuition fees and living costs ("maintenance"). Grants and some loans were means-tested, but all students could borrow the full fee, plus some extra for maintenance. The repayment schedule of the loans was made income contingent, meaning that no repayments were required until a graduate earned over a threshold amount, and repayments were set at a percentage of all earnings over this threshold. Therefore, not only does attending university affect the earnings that someone might expect to receive, but their (expected) future earnings will affect how much they expect to pay for their degree, a key feature to capture in the model.

Tuition fees, student loans and maintenance grants The funding of higher education in the UK has changed frequently in recent years moving from a model of direct government funding prior to 1998, to a model with increasingly higher tuition fees alongside a system of government-subsidised loans and grants (see Table 2.1 in Crawford and Jin (2014) for a summary of some of these changes). The majority of the individuals in the main cohort we use left sixth-form in 2007, so they would have experienced the system under reforms that came into force in 2006, henceforth the "2006 reforms". The key features of the system under the 2006 reforms are summarised in table A1.

[15]This maximum fee is set currently at slightly over £9,000, though the increase occurred after the relevant period for the analysis in this paper (in 2012).

Table A1: Details of fees, loans and grants available under the 2006 reforms

Measures	Details
Tuition fees	• Set by university, up to £3,000 p.a. • payable by ALL students
Grants	• Means-tested up to £2,700 p.a. • Tapered to zero at £33,560.
Loans	
Fees	• Equal to fees charged by university. • Available to ALL students.
Maintenance	• £3,555 p.a. if household income <£26,000. • Loan increases from £3,555 p.a. incrementally • Up to £4,405 p.a. if family income between £26,000 and £33,560. • Tapered down to £3,305 at £44,000.
Repayment	• 9% of income above £15,950 (threshold rises with inflation). • State-subsidised loans, zero-real interest rate. • Debt forgiven after 25 years.

Source: Crawford and Jin (2014)

Table A2: Expected debt on graduation (maximum loans under 2006 reforms)

Parental income	Debt on graduation	Share in sample
Low (<£15,970 p.a.)	£19,340	0.20
Middle (~£25k p.a.)	£19,340	0.09
Upper middle (~£30k p.a.)	£21,440	0.22
High (>£44k p.a.)	£18,670	0.31
Missing income info.	-	0.18

Source: Dearden et al. (2005) (debt figures) and author's calculations.

Student debt levels on graduation. Dearden et al. (2005) calculate expected debt levels for a student entering university in 2006/7 (i.e. under the first year of the 2006 reforms). Their calculated expected debts are in table A2, along with the share of individuals in each category in the Next Steps cohort. The information in tables A1 and A2 show that although the sticker price of education in the UK was quite high, loans were available to all suggesting credit constraints are not an issue in the UK context. In addition the (maximum) debt burden faced by students appears to be relatively constant across socio-economic groups (though of course the psychological effects of this debt may still vary).

1.B Data appendix

1.B.1 Next Steps

Other information collected in wave four

In addition to the data on expectations collected in wave 4, I also use information on family background and schooling up to age sixteen. I use detailed information on parental earnings to estimate a measure of socio-economic status (SES), based on the quintiles of parental earnings (I also use an alternative definition based on means-tested grant eligibility, again calculated from parental earnings). I include information on parents' occupations, ethnicity, education, and income in the model, as well as (limited) information on ability[16] (number of A-levels being taken), and gender. Table 1.2 presents descriptive statistics for these variables.

Wave eight (age twenty-five)

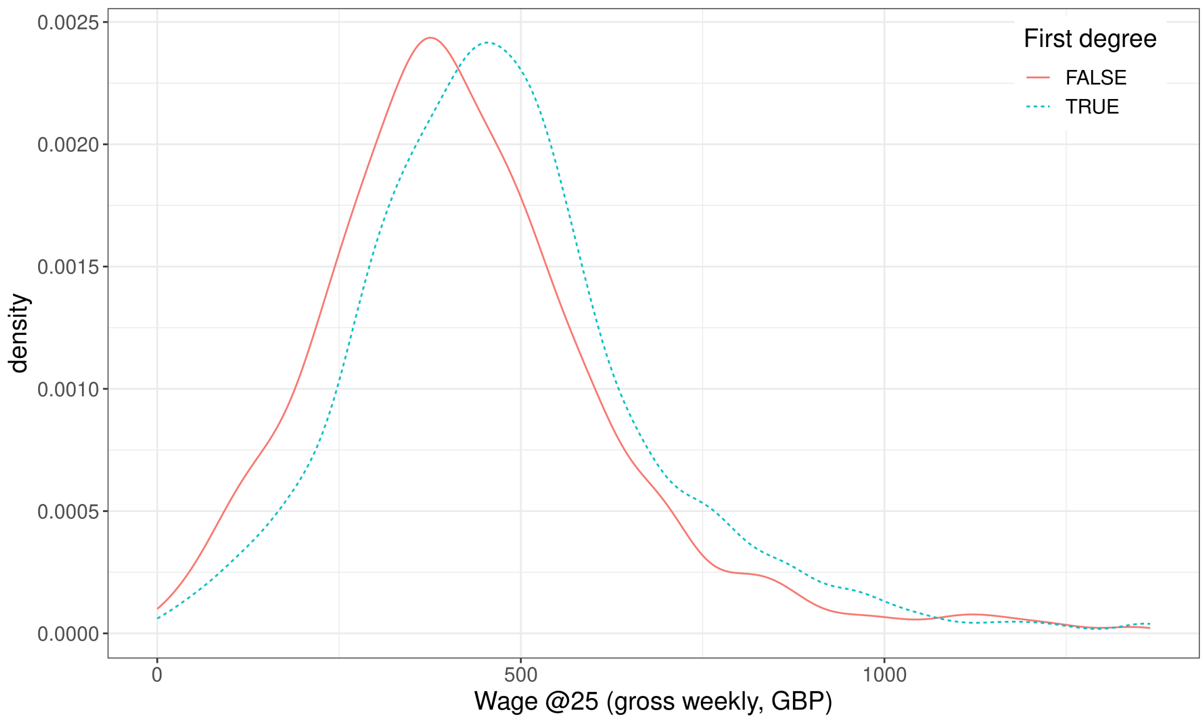
The other key wave of Next Steps for my analysis is the eighth, when the cohort members are aged 25. At this point the majority are working (or at least have worked at some point), and most of those who attend university have completed their degrees.

Degree attainment. The cohort members are asked about any qualifications they have achieved since the last interview (wave seven, five years previously), including whether they hold an undergraduate degree. Table 1.2 shows information on the proportion of cohort members who hold a degree at 25, including the proportion who attended a member of the Russell Group (a "club" of prestigious research universities in the UK). I also break down degree attainment by SES group (parental income quintiles) in table 1.2. All these statistics are shown for all respondents to waves 4 and 8, and for the subsample who answered questions about university. Nearly 70% of the analysis subsample hold a degree by the time they are 25, though there is still substantial variation across socio-economic groups reflecting the patterns highlighted in section **??**. The rate of BAs at 25 among those from the most advantaged backgrounds is 75%, compared with 60% for those from the least advantaged. That the socio-economic attainment gap persists among these "high-achieving" students suggests the issue runs deeper than performance at school.

Wages. As the majority of the cohort members are in work at age 25, a focus of wave 8 is on their careers, occupations and other labour market outcomes. In particular they are asked to provide detailed information about their wages. Figure B1 shows the distribution of weekly wages in the sample, conditional on degree attainment. The conditional

[16]The survey is linked to an administrative education dataset, the National Pupil Database (NPD), so there is much more detailed information on the students' (academic) abilities potentially available. Unfortunately, I do not currently have access to this additional data as it must be accessed in the UK and only by researchers affiliated with a UK university.

Figure B1: Distribution of weekly wages at age 25, by degree attainment



Notes: The distributions are estimated (and plotted) using the `density` option in the `R` package `ggplot2` (Wickham, 2016), using the default setting of a Gaussian kernel density estimator. Analysis subsample ($N = 4,640$).

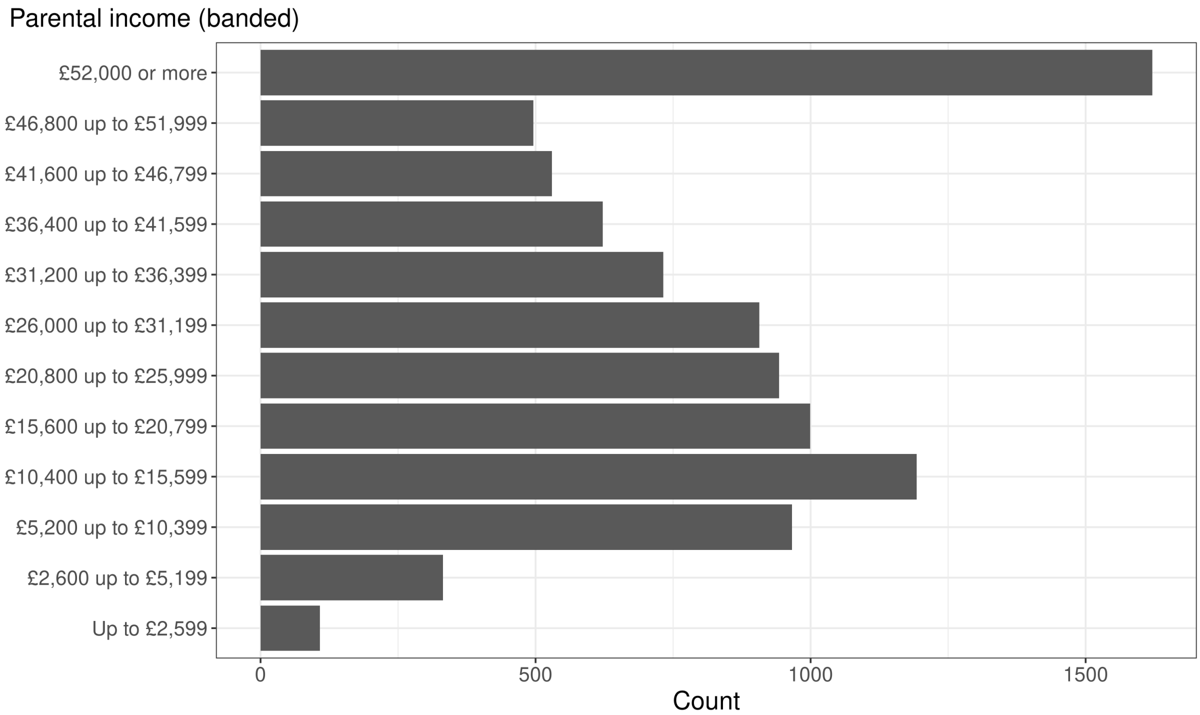
distributions look very similar, with the distribution corresponding to holders of an undergraduate (first) degree shifted slightly to the right. The mean and variance of these distributions are in table 1.2. However, such analysis does not reveal expected, nor counterfactual, wages: i.e. what graduates (expect they) would earn had they not gone to university, and vice versa. For that we need the model and assumptions detailed in section 3.2.

1.B.2 Additional information on Next Steps

Next Steps started in 2004 when the members were in secondary school aged 13 or 14. They were then interviewed annually for the next six years, until aged 18 or 19 (waves 1–7). A further round of interviews (wave 8) was conducted in 2016 when the members were aged 25 or 26, and another is planned for 2021. For consistency with the BCS data, we will focus on the data collected at age 16 (or thereabouts, wave 4) and at 25 (wave 8).

Parental income. The LSYPE records information on member's family background in waves 1–7. Though data on parental income was collected in wave 4, it was recorded in 12 bins, with the top (and most populous) bin starting at £52,000 p.a. (see figure B2). More detail was collected in wave 1—over 30 bins, plus further information for some top-coded families—as well as continuous data on parents' salaries for some families (see figures B3 and B4).

Figure B2: Parental income in the LSYPE (wave 4)

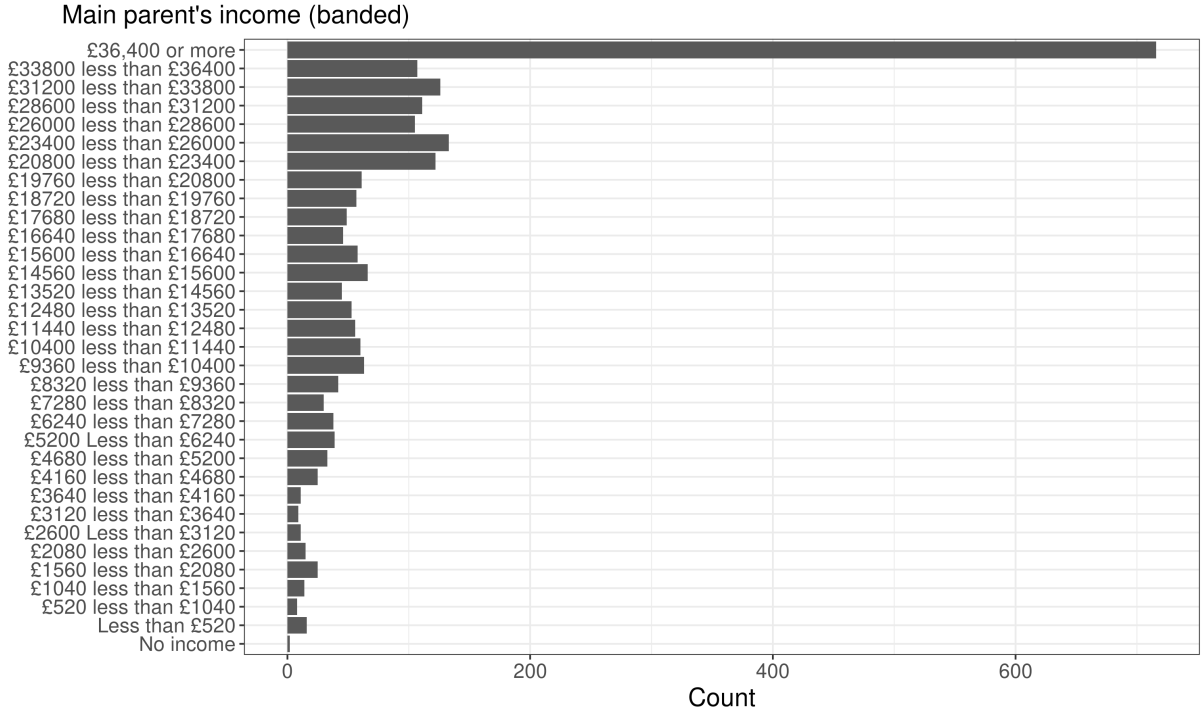


Source: LSYPE wave 4 (CLS, 2018).

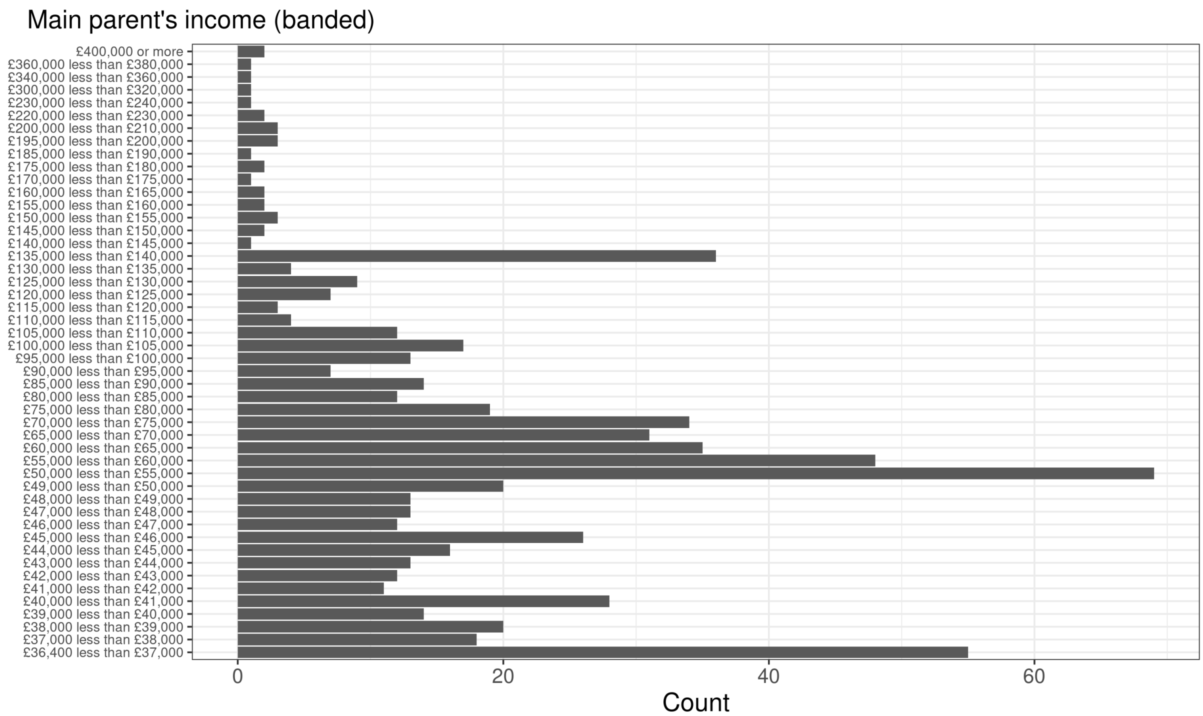
Undergraduate degree. Figure B5 shows the proportion of individuals in the LSYPE who hold a degree at 25, broken down by gender.

Figure B3: Main parent's income in the LSYPE (wave 1)

(a) Banded



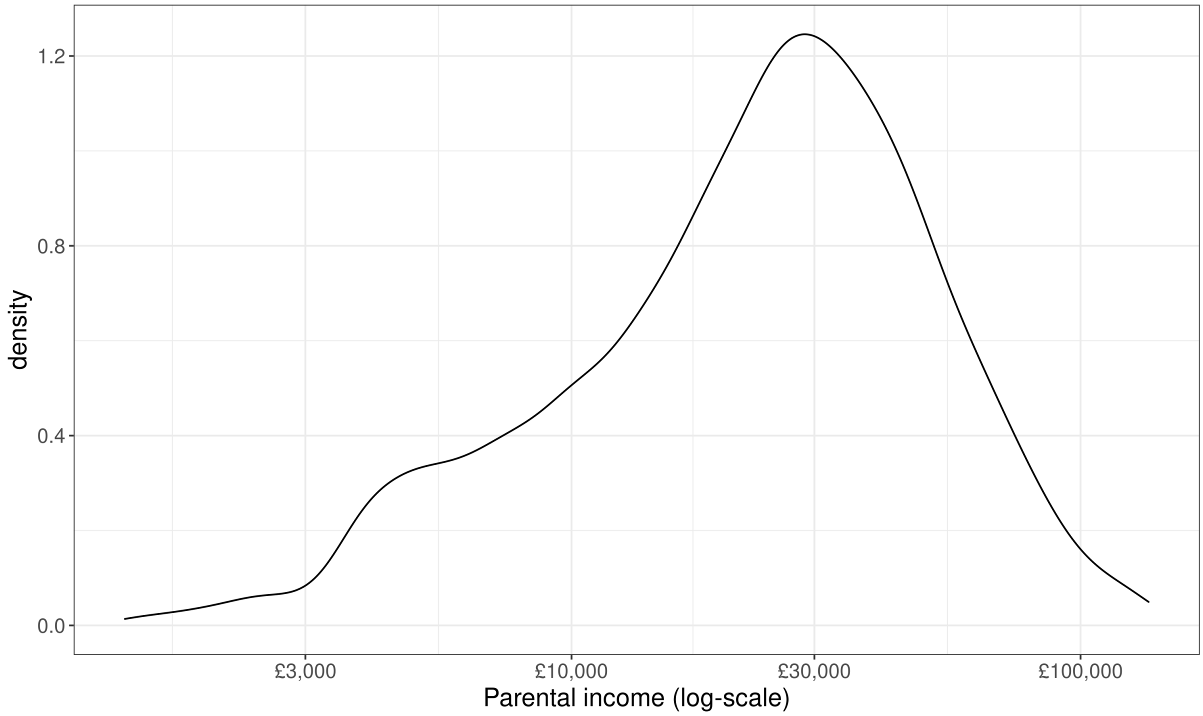
(b) Top-code ($> £36,400$) detail



Source: LSYPE wave 1 (CLS, 2018).

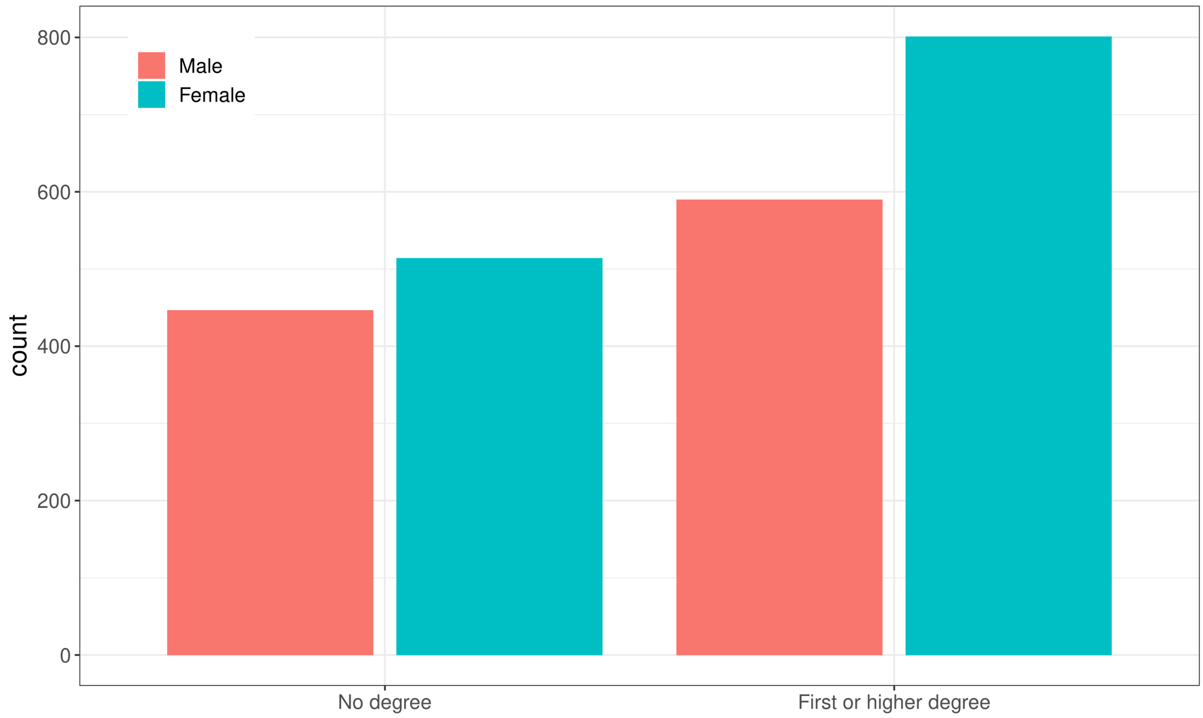
Notes: The top panel (a) shows all recorded earnings for main parents in wave 1. Panel (b) shows a detailed breakdown of the top band from panel (a).

Figure B4: Parental income in the LSYPE (wave 1, density)



Source: LSYPE wave 1 (CLS, 2018).
Notes: This plot shows the density of (log-)annual earnings, calculated using the default kernel density estimator of the `geom_density()` function in the `ggplot2` `R` package (Wickham, 2016).

Figure B5: Undergraduate degree at 25 by gender (LSYPE)



Source: LSYPE wave 8 (CLS, 2018).

Table B1: The advantages (+) and disadvantages (−) of going to university

Response (harmonised)	+/−
Career	
Will lead to a good/better job (than would otherwise get)	+
Will lead to a well paid job	+
Gives someone better opportunities in life	+
Is essential for the career they want to go into	+
Shows that you have certain skills	+
To delay entering work/ more time to decide on a career	+
Not being able to start earning money/start work	−
No guarantee of a good job at the end	−
Don't need to go to university for the job someone may want	−
Get less work experience	−
Financial / debt	
Now	
It is expensive	−
Not becoming financially independent	−
Not being able to start earning money/start work	−
Costs (general/non specific)	−
Tuition fees/Accommodation costs/Living expenses	−
Future	
Will lead to a well paid job	+
Getting into debt/have to borrow money	−
Social life / environment	
The social life/ lifestyle / meeting new people / it's fun	+
To leave home/ get away from the area	+
Leaving home/family/friends	−
Stress	−
Education	
To carry on learning / I am good at / interested in my chosen subject	+
Get more qualifications/better/higher qualifications	+
The workload can be hard/ doubts about ability to finish course	−
Personal development	
Makes someone independent/ maturity / personal development	+
Gives you more confidence	+
People will respect me more	+
Leads to a better life/good life (general)	+
Prepare you for life/gain life skills	+
Time	
To delay entering work/ more time to decide on a career	+
Takes a long time	−
Waste of time (general/non-specific)	−

Table B2: Logit estimates (all responses, first-stage)

(a) Advantages

Dependent variable:	Degree
Get better job	0.422***
	(0.099)
Well-paid job	0.266***
	(0.103)
Better opportunities	0.454***
	(0.100)
Need for career	−0.028
	(0.226)
Show skills	0.194
	(0.299)
Delay get job	0.670
	(0.569)
Social life	0.138
	(0.116)
Leave home	0.052
	(0.287)
Learning	0.297**
	(0.149)
More qualifications	0.055
	(0.096)
Personal development	0.621***
	(0.158)
More confidence	1.109**
	(0.513)
More respect	−0.135
	(0.630)
Better life (general)	0.134
	(0.272)
Gain life skills	0.298*
	(0.154)
Other	0.208
	(0.232)
Don't know	−0.012
	(0.296)
No answer	−0.719*
	(0.373)
Observations	3,469
Log Likelihood	−1,985.799
Akaike Inf. Crit.	4,105.599

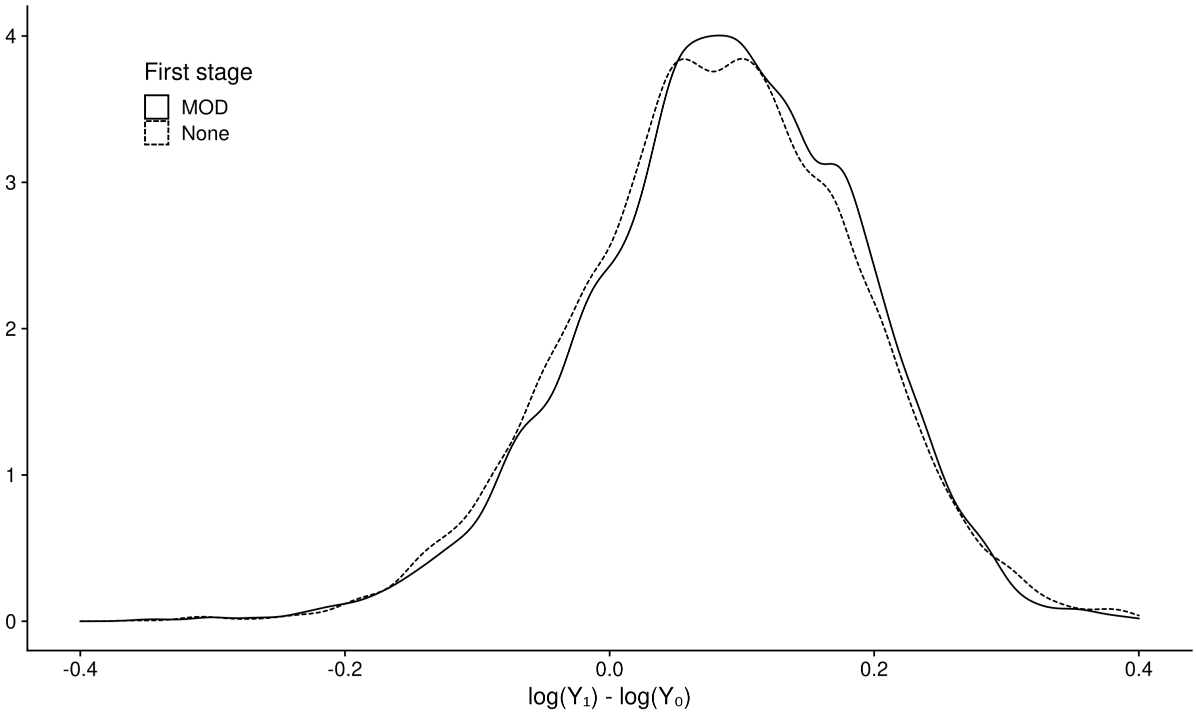
(b) Disadvantages

Dependent variable:	Degree
Expensive	0.108
	(0.152)
Get into debt	0.141
	(0.112)
Depend on parents	12.457
	(228.304)
Not financially indep.	−0.013
	(0.385)
Not earning / working	−0.092
	(0.169)
Costs (general)	0.200*
	(0.121)
No job guarantee	0.065
	(0.162)
Not needed for job	−1.163**
	(0.544)
Less experience	−0.223
	(0.278)
Heavy workload	0.256
	(0.175)
Leave home	−0.286*
	(0.151)
Takes long time	−0.273*
	(0.155)
Waste of time	−0.272
	(0.369)
Tuition fees etc.	0.400*
	(0.229)
Stress	0.523
	(0.372)
Other	−0.197
	(0.159)
Don't know	−0.134
	(0.244)
No answer	−0.100
	(0.190)
Observations	3,469
Log Likelihood	−1,985.799
Akaike Inf. Crit.	4,105.599

Notes: *p<0.1; **p<0.05; ***p<0.01. The two panels contain estimates from the same regression, which also included the following background characteristics: ethnicity, gender, A-levels, parental income, and a self-assessed ability measure.

1.C Results

Figure C1: Comparing wage premium distributions with and without selection correction



Chapter 2

Revisiting the returns to higher education: heterogeneity by cognitive and non-cognitive abilities

Abstract

Recent work has highlighted the significant variation in returns to higher education across individuals. We develop a novel methodology—exploiting recent advances in the identification of mixture models—which groups individuals according to their prior ability and estimates the wage returns to a university degree by group. We prove the non-parametric identification of our model. Applying our method to data from a UK cohort study, our findings reflect recent evidence that skills and ability are multidimensional. Our flexible model allows the returns to university to vary across the (multi-dimensional) ability distribution, a flexibility missing from commonly used additive models, but which we show is empirically important. The returns to higher education are 3–4 times larger than the returns to prior cognitive and non-cognitive abilities. Returns are generally increasing in ability for both men and women, but vary non-monotonically across the ability distribution.

2.1 Introduction

Economists have long concerned themselves with estimating the wage returns to education, beginning (at least) with Mincer in 1958. For almost as long, critiques of this work have argued that a failure to control for ability has led to a significant "ability bias" in estimates of the returns to education. These claims, whether well-founded or not (Griliches, 1977), led to attempts to circumvent the ability bias problem using instrumental variables. However, these estimates were *higher* than the ability-biased estimates, which were themselves supposed to be biased upwards. This prompted a new approach, recognising that the returns to education are not homogeneous, and so using different methods and instruments would lead to different estimates (Blundell et al., 2005).

By allowing heterogeneous returns to higher education, we join this growing literature which explicitly studies the variation in returns to a university degree across across individuals and groups of individuals. The variation in returns across different groups can be considerable.[1] However, there is little evidence on how the returns to education vary with a fundamental characteristic, ability.[2] Ability is important not only as a potential source of bias; lessons from the literature on human capital formation[3] suggest a person's ability is also likely to directly impact the returns they achieve from a university degree. This paper develops and estimates a framework explicitly designed to investigate how the returns to university vary flexibly with what we call "prior ability": a young person's cognitive and non-cognitive ability on entry to university.

Our focus on cognitive and non-cognitive ability recognises the growing body of evidence that skills are multidimensional, and that collapsing these dimensions into a single (usually cognitive) measure misses important sources of variation across people.[4] Our

[1]Britton et al. (2021a) investigate how the returns to university vary across socio-economic and ethnic groups in the UK. They find positive returns to university for all groups, though substantial heterogeneity: returns are higher for women than for men, and across ethnic groups they vary from 7% for White British men, to 40% for Pakistani women. Britton et al. (2021b) study the returns to different subjects and institutions, again finding substantial heterogeneity in the returns to different subject and institutions after controlling for prior cognitive ability. They find weak evidence that returns are positively correlated with the selectivity of the subject or institution.

[2]We use the terms "human capital", "skills", and "ability" interchangeably throughout this paper.

[3]Recent work by Cunha, Heckman and coauthors (2006; 2007a; 2008; 2010) has shown that skills obtained early in childhood are vital for fostering skills later in childhood—a feature they call the *self-productivity* of skills. A related concept is the *complementarity* of skill formation: "skills produced at one stage raise the productivity of investment [in further skills] at subsequent stages" (Cunha et al., 2006, p. 703). These features of skill production during childhood suggest that ability on entry to university will affect the impact of a university degree on an individual's ability, and hence on their later outcomes.

[4]Focusing on educational outcomes, Jacob (2002) finds that non-cognitive skills are key in explaining the gender gap in college attainment in the US, and Delaney et al. (2013) demonstrate the link between non-cognitive skills and study behaviours known to be important for success in undergraduate degrees. Turning to success later in life, Heckman et al. (2006) offer evidence that non-cognitive skills are important for a range of social and economic outcomes. Bowles et al. (2001) survey the literature on the determinants of earnings, with a particular focus on noncognitive traits. More recently, Todd and Zhang (2020) include personality traits in a dynamic discrete choice model of schooling and occupational choice. The authors find important links between personality and schooling, and between personality and occupational choice.

analysis incorporates these insights by allowing both cognitive and non-cognitive skills to determine wages, and hence the wage returns to a university degree. We compare results obtained using only cognitive measures with those using both cognitive and non-cognitive measures, thereby providing further evidence of the important role for non-cognitive ability.

A key contribution of our paper is methodological: we adapt the framework of Cassagneau-Francis et al. (2021), where we proved the non-parametric identification of, and developed an estimation strategy for, a model of formal training and wages.[5] This work continues a long tradition in economics of using discrete mixtures to model heterogeneity, going back to Heckman and Singer (1984). In recent years, major progress has been made in the identification of this type of model, and in the development of nonparametric estimators (see for example Bonhomme et al., 2016a and the references therein). Here and in Cassagneau-Francis et al. (2021) we make novel attempts at using discrete mixtures in the context of evaluation models.

Our statistical model, motivated by the human capital formation literature mentioned above, builds upon the work of Heckman and coauthors (2003; 2006; 2007a; 2010). In these papers, the analysis typically requires strong functional form assumptions to identify the model, often assuming an underlying factor model structure (Carneiro et al., 2003).[6] By adapting the framework in Cassagneau-Francis et al. (2021), we are able to achieve identification of a yet more flexible model, estimate a non-linear version of our model, and demonstrate the importance of these non-linearities empirically.[7]

In order to identify our non-linear model, we assume that the distribution of prior ability (i.e. of latent types in our framework) is discrete. Our method is frugal in its requirements of the data available to the econometrician, and having discrete types means our heterogeneity analysis is easily interpreted. The costs of this flexibility, frugality, and interpretability are low: (i) we require, in addition to a measurement of each component of ability and an outcome,[8] a single crude (i.e. discrete) measurement of ability, *or* a discrete instrument for university attendance (though exogeneity is only required conditional on type), and (ii) we assume the distribution of prior ability has finite support.

Following Cassagneau-Francis et al. (2021), we show the non-parametric identification of the returns to university conditional on an individual's prior ability, which in our model is summarised by their latent type. We can aggregate these type-conditional "treatment

[5]Similar techniques have been used to identify a range of models including: firm and worker sorting (Bonhomme et al., 2019); and the contributions of workers across different teams (Bonhomme, 2021). Gary-Bobo et al. (2016) identify and estimate a parametric model, also inspired by the human capital formation literature, to study the effects of grade retention on French middle school students.

[6]Cunha et al. (2010) show how to relax some of the stronger assumptions, allowing the measurement and outcome equations to be non-linear in their inputs.

[7]Cunha et al. (2010) only estimate the additive version of their model.

[8]Cunha and Heckman and coauthors require at least two measurements per component, plus an outcome.

effects" (TE) to obtain more standard effects: average treatment effect (ATE), average treatment on the treated (ATT), and Imbens and Angrist (1994)'s local average treatment effect (LATE). We can also aggregate the type conditional TEs to obtain analogues of the usual OLS and IV estimates, which we show to be biased estimators of ATT (for OLS) and LATE (for IV). Having shown non-parametric identification, we adapt our estimation strategy from Cassagneau-Francis et al. (2021), specifying a parametric specification of our model. This approach allows us to use a sequential version of Dempster et al. (1977)'s expectation-maximisation (EM) algorithm. We use a bootstrap procedure to calculate standard errors.

In our application, we use our framework to estimate the returns to a university degree in the UK as a function of cognitive *and* non-cognitive prior ability. Our data come from the British Cohort Study (BCS 1970), which follows all individuals born in the UK in a single week in 1970, and contains detailed information on the cohort members at age 16 (before attending university) and again at age 26 (after university). In particular, the young people took cognitive tests and answered a series of questions to capture their non-cognitive abilities at age 16. Crucially we also observe their wages at age 26, along with any qualifications they have achieved up to that point and hence whether they graduated from university.

We estimate our model separately by gender,[9] a data-driven decision resulting in better performance from our estimation algorithm. Although the graduation rates for men and women are quite similar (33% for men, and 27% for women), there was a large gender wage gap during this period (reflected in our data), both for graduates and non-graduates. Our algorithm struggled to deal with this difference when we pooled genders.[10] A large and important literature attempts to uncover the institutional and societal factors driving this gap, a task which is beyond the scope of this paper.

We find that the returns to a university degree for our UK cohort are generally positive and large for both men (10–20%) and women (15–28%), but vary significantly with prior ability—i.e. across individuals of different types in our framework. This variation is also highly non-linear, with the size of the effect varying non-monotonically across the ability distribution. However, these patterns are quite different across genders. For men, the returns are U-shaped with respect to prior ability, with middle-ability types receiving the lowest returns. The opposite is true for women, for whom we observe hump-shaped returns with the highest for middle-ability types.

This non-linearity would not be apparent under the current leading estimation approaches which assume an additive model for wages. In a linear (additive) model, the

[9]Blundell et al. (2000) also estimate their model separately by gender, and study a similar UK cohort born 12 years earlier than our cohort, though consider wages at 33, so 5 years earlier than we observe wages.

[10]Estimating our model separately across genders is the most straightforward way to "control for" gender in our framework.

wage returns to university are proportional to a young person's ability level. Therefore, if the returns to university are increasing in ability *on average*, we would estimate a higher return for a high-ability young person than a low-ability young person—even if this relationship only holds for part of the ability distribution.

Our analysis also reveals that the returns to a university degree in the UK are more important than the returns to ability in the following sense: a low-ability young person can earn higher wages by completing university, and becoming a low-prior-ability graduate, than they could by improving their ability to become a high-ability non-graduate. The large impact of university on wages across the ability distribution drives another of our main results: the contribution of the graduate wage premium to wage inequality is 3 (men) and 4 (women) times larger than the contribution of ability attained *prior to university*. Our results complement those of Cawley et al. (2001) who find that, having controlled for educational attainment, cognitive ability explains little of the variation in wages across individuals even within occupations. We find that *both* cognitive and non-cognitive skills explain only a small part of wage inequality.

Despite the relatively small *direct* contribution of ability to wage inequality, a young *man's* levels of cognitive and non-cognitive skills on entry to university do influence the returns they can expect to achieve. There is a significant comparative advantage for non-cognitive skills among male non-graduates, resulting in low returns to university for high non-cognitive-ability, middle-cognitive-ability men. The equivalent is not true for women. To what extent this is due to the different occupations favoured by male and female non-graduates remains a question for future research.

The remainder of the paper proceeds as follows. In section 3.2 we present the setup of our model, with a discussion of identification in 2.2.1. Section 2.3 describes how we estimate the model. We then turn to our application using UK-cohort data: estimating the wage returns to a university degree as a function of cognitive and non-cognitive prior ability. Section 2.4 discusses the relevant context of higher education in the UK, and presents our dataset and some initial descriptive results. Section 2.5 presents the results of estimating our model on this data, first with only cognitive ability, and then with both cognitive and non-cognitive components. Section 2.6 concludes.

2.1.1 Related literature

Returns to education. There is a long tradition in economics of attempts to estimate the returns to schooling, a tradition which perhaps began with the seminal work of Mincer (1958, 1974). Critiques of this early work suggested it was plagued by issues of ability bias, which although arguably small (Griliches, 1977), led to a search for sources of exogeneous variation and the use of IV methods to avoid this criticism. Card (1999) provides an excellent summary. However, despite these methods being used to avoid the

positive ability bias, IV estimates of the returns to schooling are typically larger than OLS estimates. These apparently contradictory findings were due to either an even larger ability bias, for family background IVs, or particularly high *marginal* returns for those impacted by institutional IVs — Imbens and Angrist (1994)'s local average treatment effect, or LATE.

These high marginal returns estimated by IV methods highlighted another avenue to explore: the returns to education are unlikely to be constant, varying with both observed and unobserved characteristics. Initial work used sibling Altonji and Dunn (1996) and twin (Ashenfelter and Rouse, 1998) studies to analyse the effects of family background on the returns to education, finding little variation. Barrow and Rouse (2005), also focusing on siblings, find little effect of race and ethnicity on the returns to education.

Much of this work focused on the return to an additional year of schooling. Other authors have focused on the returns to educational milestones, with the returns to a university degree being most relevant for this paper (see Kane and Rouse (1995) for evidence from the US and Blundell et al. (2000) from the UK). Allowing returns to be heterogenous, Carneiro et al. (2011) estimate returns to college that vary with the unobserved cost of attaining a degree. A recent paper by Britton et al. (2021b) studies how the returns vary across different degrees and institutions, as well as across different socio-economic and ethnic groups. We join this growing literature on the heterogeneous returns to a university degree, estimating wage returns which vary with both cognitive and non-cognitive ability.

Another relevant strand of the literature compares the returns to ability with those to education. Taber (2001) argues that the growth in the wage premium in the US in the 1980s is largely driven by an increase in the demand for high-skill (i.e. college-education) workers. Cawley et al. (2001) find that cognitive ability explains only a small part of wages once schooling is controlled for, and highlight that non-cognitive ability is also important for labour marker outcomes.

Human capital formation and non-cognitive ability. The model in our paper is inspired by the literature on human capital formation. This literature, which mainly focuses on the production of human during childhood, contains a number of lessons relevant for our analysis. Early work on human capital distinguished it from "ability", as being something that something that could be invested in, unlike innate ability which was invariant (Becker, 1964). The focus was entirely on cognitive ability (and human capital). More recent work has argued that there is no true distinction between human capital and ability — whether called skills, ability, or human capital, these traits are a product of an individual's genes, environment, and can be acquired and improved; and that both cognitive and non-cognitive abilities are important, both for success during formative years by fostering further improvements in these abilities, and also for later outcomes. These

findings are summarised by Cunha et al. (2006, henceforth CHLM) in an excellent review.

We borrow a number of insights from this literature. CHLM emphasise two key related features of skill formation which we also incorporate into our analysis: (i) skills produced at one stage of development are important for fostering skills at later stages, CHLM's "self-productivity" of skills; (ii) later investment in skills is necessary to fully realise the benefits of earlier investments — in CHLM's terminology the "complementarity" of skills. Our contribution is to embed these insights from childhood development into a model of investment in skills at a later stage of the life-cycle: higher education. We design a framework to study the returns to higher education, explicitly allowing returns to vary with prior cognitive and non-cognitive abilities. As far as we are aware, our paper is the first to estimate a model of this type allowing for non-linear measurements and outcomes.

Model, identification and estimation. Our empirical framework is close in spirit to the recent papers of Cunha and Heckman (2007a, 2008) and Cunha et al. (2010) who aim is to estimate the technology of skill formation. Similar to the aforementioned papers, we assume latent factors which link measurements and outcomes, and like Cunha et al. (2010) we are able demonstrate the non-parametric identification of our model.

We depart from this work in assuming a discrete distribution for these latent types. Similar assumptions have recently been used to model unobserved heterogeneity by a number of authors. Bonhomme and Manresa (2015) investigate the properties of using latent types (or groups) to capture (unobserved) heterogeneity, in a setting the authors call "group fixed effects". This type of setup is explored further in Bonhomme et al. (2022). These methods have been applied to study matched employer-employee data (Bonhomme et al., 2019) and the contributions of individuals in team settings (Bonhomme, 2021).

Closest to our work are the papers by Gary-Bobo et al. (2016) and Cassagneau-Francis et al. (2021). Gary-Bobo et al. (2016) estimate the effects of grade retention on French middle school students, while Cassagneau-Francis et al. (2021) estimate the wage returns to formal training in France, both relying on discrete types to capture unobserved heterogeneity. Both these papers employ on a differences-in-differences-like setup, observing the same outcome before and after treatment — our paper adapts this method to allow different measurements/outcomes before and after treatment.

We follow the identification proof in Cassagneau-Francis et al. (2021), which relies on recent advances in the identification of finite mixtures, using matrix algebra to prove identification. We refer the interested reader to a series of papers by Bonhomme et al. (2016a,b, 2017) and citations therein, for further details. We use Dempster et al. (1977)'s EM algorithm to estimate our model, following both Gary-Bobo et al. (2016) and Cassagneau-Francis et al. (2021). This method has been used widely in economics, providing a relatively straightforward method to estimate models with "missing" data (our latent types). However, it can still be computationally intensive, involving maximising complex likeli-

hood functions. Arcidiacono and Jones (2003) show how an alternative formulation of the problem allows *sequential estimation*, allowing parameters to be updated separately.

2.2 Empirical framework

There are N young people indexed by i. We denote their (log)-wage by w_i, observed at age 26 when they are either university graduates, denoted $d_i = 0$, or non-graduates $(d_i = 0)$. Their wage depends on their ability before attending university, which we will call "prior ability" and denote by θ, and on whether they graduate from university. Ability is multi-dimensional. We focus on the two-dimensional case, in which individuals might differ in their cognitive (θ^C) and non-cognitive (θ^N) abilities. Then $\theta = (\theta^C, \theta^{NC})$. The different components of ability may or may not be correlated. Our aim is to estimate the causal impact of graduating from university on wages, as a function of prior ability.

However, we do not observe ability (θ) directly. Ability is the classic confounder in attempts to estimate the returns to university, being both a determinant of wages *and* of the decision to attend university (Becker, 1964; Card, 1999). It is also more fundamental to our analysis, given we want to study how the returns to university vary with prior ability. We follow the example of both the recent literature on human capital formation and on the returns to schooling (see for example Cunha and Heckman, 2007a; Carneiro et al., 2011), relying upon noisy measurements of a young person's prior ability. We have (at least) one measurement specific to each ability, i.e. a purely cognitive measurement and a purely non-cognitive measurement. Using the information on θ contained in these measurements and in wages, we are able to identify and estimate the distribution of θ, and hence study how the returns to university vary across this distribution.

We depart from this recent literature in how we model the dependence of measurements and wages on ability. Typically, authors assume mean measurements and wages are linear functions of ability, with higher moments of the distribution independent of ability (see for example Carneiro et al., 2003). A linear version of our model is presented and briefly discussed in appendix 2.A. We relax these assumptions, and imposing no functional form on how the means of these distributions depend on ability, and we can allow the variance to be a function of ability (though we do not in our application). To achieve this, we assume the distribution of prior ability has finite support. Under this assumption, we can classify individuals into a finite number of groups based on their prior ability, groups across which the distributions of measurements and wages vary systematically. We denote these groups by $k \in \{1, \ldots, K\}$. Therefore, ability takes only a finite number of values, which we can index by the group identifier, k, so that $\theta_k = (\theta_k^C, \theta_k^{NC})$.

In addition to the continuous wages and measurements there is a discrete variable z, which is either an additional (crude) measurement of ability, or an "instrument" for university graduation. We do not include any other control variables when discussing

identification, or when estimating our model. However, adapting our proof and estimation strategy to include controls would be straightforward, though it would require additional restrictions on our model. We say "instrument" as z need only be independent of measurements and wages *conditional on prior ability*, i.e. on a young person's type. This idea is formalised in our first assumption.

Assumption 1 (Measurements and wages)**.** *Measurements, wages and z are independent conditional on type and education.*[11]

We denote the distribution of wages conditional on type and education by $f_w(w_i|k,d)$. Similarly, the conditional distribution of the measurements is $f_\ell(M_i^\ell|k,d)$, $\forall \ell \in \{C, NC\}$. The probability mass of young people of type $k \in \{1,\ldots,K\}$, with value of the instrument $z \in \{1,\ldots,Z\}$, and with education level $d \in \{0,1\}$, is denoted by $\pi(k,z,d)$.[12] We want to identify and estimate these objects, along with the distribution of prior ability ($\theta_k, \forall k = 1,...,K$). Before discussing our identification strategy, we present briefly the economic foundations of our statistical model.

Economic motivation for the model. Our statistical model is motivated by the literature on human capital formation. Consider a simple model of human capital formation (Todd and Wolpin, 2003; Cunha et al., 2006). There are two periods: before university ($t = 0$, age sixteen in our application); and after university ($t = 1$, age twenty-six in our application). Human capital (ability) in period $t+1$ is a function of human capital in the previous period, θ_t and any investments in human capital made between t and $t+1$, I_t.

$$\theta_{t+1} = f_\theta(\theta_t, I_t) \tag{2.1}$$

We can simplify the notation in equation (2.1) to include just this single period, between $t = 0$ and $t = 1$, and we additionally assume that investment in human capital during this period is binary: young people either attend and graduate from university or they do not. We can also use k and θ_0 interchangeably, so

$$\theta_1 = g_\theta(\theta_0, d) = g(k,d)$$

[11]Measurements need not be independent of each other even conditional on type.

[12]Our model is closely related to the extended Roy model (Heckman and Vytlacil, 2005; Carneiro et al., 2010, 2011):

$$w = w(k,0) + [w(k,1) - w(k,0)]\, D$$
$$D = 1 \text{ if } \mathbb{E}[w(1) - w(0)|k] \geq c(k,z),$$

where k is an individual's type (capturing their cognitive and noncognitive ability). z is the instrument, i.e. an environmental variable affecting treatment decision, through the non-pecuniary cost of attending university, but independent of wages and measurements conditional on type. $w(k,0), w(k,1)$ are treatment-specific outcome variables (random given k and independent of z). $c(k,z)$ is cost of attending university (random given k,z).

Then, if wages in $t=1$ are a function of human capital in $t=1$, we obtain our model for wages

$$w_i \sim \tilde{f}_w(w|\theta_1) = \tilde{f}_w(w|g(k,d)) = f_w(w|k,d).$$

2.2.1 Non-parametric identification

One of the key contributions of this paper is a novel identification and estimation strategy that does not rely on wages and measurements being linear in their components. Our strategy requires fewer measurements of prior ability for identification than the current leading factor-model approach, and we are able to identify and estimate a fully non-linear model.[13] This frugality and flexibility come at a low cost, as compared with the linear factor-model approach we additionally require: 1) a crude (can be binary or discrete) measurement of prior ability,[14] *or* a crude instrument for university attendance, and which need only be exogenous of wages and measurements conditional on prior ability, which we denote z; 2) that the distribution of prior ability has finite support.

We assume during our discussion of identification that the econometrician knows the *true* number of points of support, K. However, we also discuss how this can be estimated when we operationalise the method. Recall that our aim is to identify the discrete distribution of prior ability, θ_k and $\pi(k)$,[15] the distributions of measurements and wages conditional on prior ability and education, $f_\ell(M^\ell|k,d)$ and $f_w(w|k,d)$.

Individual measurements and wages are not directly informative of prior ability due to noise in these measures / outcomes. However, under our maintained assumptions, mean wages and measurements across individuals of the same type—who share the same prior ability, θ—*are* informative. Therefore we can use these means to identify the support of θ, i.e. θ_k, and to set it in an interpretable metric allowing comparisons *across* types.

Likelihood of an individual's observations. Under the model detailed at the start of section 3.2, the likelihood associated with an individual i's observations writes

$$\ell(\boldsymbol{M}_i, w_i, d_i, z_i) = \sum_k \pi(k, z_i, d_i) f_M(\boldsymbol{M}_i|k) f_w(w_i|k,d) \tag{2.2}$$

with $\boldsymbol{M} = (M^1, \ldots, M^L)$, $f_M(\boldsymbol{M}|k) = \prod_\ell f_\ell(M^\ell|k)$.

We follow Cassagneau-Francis et al. (2021) and apply results from recent work on mixture models to show the elements on the right-hand side of equation (2.10) are identified

[13] Cunha et al. (2010) prove identification of a non-linear version of the factor model, but rely on additively separable measurement and outcome equations when estimating their model. Their method requires the same number of observations (measurements) as the linear model.

[14] This measurement is not really *additional* to the factor model; we could use one of the extra measurements required in that approach, discretising the variable if it is continuous.

[15] Our method identifies $\pi(k,z,d)$ which we can then sum over z and d to obtain $\pi(k)$, the proportion of individuals of type k.

under certain conditions (see Bonhomme et al. (2017) for details). A formal statement of the necessary assumptions, our identification theorem and a detailed proof is in appendix 2.B. Here, we only summarise the key assumptions and ideas of the proof.

We have already introduced one of the main assumptions (assumption 1); that measurements, wages and the instrument (if used) are independent conditional on type. This allows for dependence of higher moments of the measurement and wage distributions on latent types (i.e. on ability), not just the means of these distributions.[16] The key here is that all the dependence across wages and measurements is summarised by a person's type.

Next, we require the wage and measurement distributions (excluding z) to be continuous, or at least sufficiently granular that the type-conditional wage and measurement distributions are linearly independent. We cannot identify any latent types whose wage (measurement) distribution is a linear combination of the wage (measurement) distributions of the other types.

There are then two conditions on the probability mass function, $\pi(k,z,d)$. The role of z in the identification is to form similar systems, one for each value of z. These systems are similar in that they contain the same measurement and wage distributions, but we rely on z to ensure they are sufficiently different to allow identification of all the components.[17] For this, z either needs to be correlated with d, but not with wages (conditional on k) i.e. z is an "instrument"; or z needs to be correlated with k, i.e. z is an additional measurement; or both.

Finally, we need to label types consistently across values of d. Our method identifies $f(M|k,d)$, $f_w(w|k,d)$, and $\pi(k,z,d)$ separately for each value of d. But, there are no young people with all values of d that would allow us to label types consistently. Therefore we need another assumption. We assume that the measurement distributions are independent of education, conditional on type. Then, we can equate these distributions across values of d to label k consistently.

Having identified these distributions, we are able to calculate heterogeneous returns to university, conditional on prior ability. We are in a *potential outcomes* framework, with each individual having two potential outcomes, w_1 and w_0, of which we only observe one. The ATE under this framework is $\mathbb{E}[w_1 - w_0]$, although only $\mathbb{E}[w_1 \,|\, d=1]$ and $\mathbb{E}[w_0 \,|\, d=0]$ are observed. If we believe that wages are correlated with education, then

$$\mathbb{E}[w_1 - w_0] \neq \mathbb{E}[w_1 \,|\, d=1] - \mathbb{E}[w_0 \,|\, d=0]. \tag{2.3}$$

[16]This might be important as one can imagine some young people being particularly good (or bad) at tests, a "skill" that would affect both their cognitive and non-cognitive scores, but one that might not neccessarily be valued by employers.

[17]We refer the reader to the proof in appendix 2.B for formal notions of similar and different in this context.

This is the classic problem of *selection into treatment*, where treatment in our case is a university degree.

Fortunately, our maintained assumptions permit a solution. Potential outcomes are independent of education and z conditional on prior ability, i.e. $w_1, w_0 \perp\!\!\!\perp z, d \,|\, k$. We can then define the following type-conditional "treatment effect".

Average treatment effect by type, $ATE(k)$. This is the expected wage gain from a university degree for a young person of type k, and perhaps *the* key object of our analysis:

$$\begin{aligned} ATE(k) \equiv \mathbb{E}[w_1 - w_0 \,|\, k] &= \mathbb{E}[w_1 \,|\, k] - \mathbb{E}[w_0 \,|\, k] \\ &= \mu_1(k) - \mu_0(k). \end{aligned}$$

In appendix 2.C, we show how to aggregate these type-conditional returns to obtain the usual estimands that analysts estimate when considering the returns to education—average treatment effect (ATE), average treatment on the treated (ATT)—within our framework. We also demonstrate some of the biases from which these standard estimation approaches suffer.

The linear factor-model approach that the current state-of-the-art allows one to study the heterogeneity in returns to education, and to compare the contribution of education to wage dispersion with the contribution of prior ability. One can also study the correlation between cognitive and noncognitive abilities. However, the linearity assumption shuts down any interaction (e.g. complementarities) between the different components of prior ability both on wages directly, and on the returns to education. This approach also requires homoscedasticity: error terms cannot depend on the level of prior ability. We are able to relax this assumption for both the wage and measurement equations.

2.3 Estimation strategy

Although the non-parametric identification proof detailed in appendix 2.B is constructive and hence suggests a method to operationalise our framework, we prefer an alternative semi-parametric approach via the EM algorithm. In particular, this avoids the necessity of discretizing the measurement and outcome distributions, allowing us to use all the available information in these observations.

Following our approach in Cassagneau-Francis et al. (2021), we assume that the measurements are normally distributed conditional on prior ability, and that log-wages are normally distributed conditional on prior ability and education. Therefore, measurement M_j has probability density function (PDF)

$$f_j(M_j|k) = \phi\left(\frac{M_j - \alpha_j(k)}{\omega_j(k)}\right),$$

where $\phi(\cdot)$ is the standard normal PDF.

Similarly, log-wages, w, are distributed as[18]

$$f(w|k,d) = \frac{1}{\exp w}\phi\left(\frac{w-\mu(k,d)}{\sigma(d)}\right).$$

2.3.1 EM algorithm

Having made parametric assumptions, we can now use Dempster et al. (1977)'s expectation-maximisation (EM) algorithm to estimate the parameters of the model via maximum likelihood (ML). The computational burden can be further reduced by applying Arcidiacono and Jones (2003)'s sequential-EM algorithm which avoids having to estimate many parameters in one step.

The ML estimator of the parameters, $\Omega = \{\pi(z,d|k), \alpha_j(k), \omega_j(k), \mu(k,d), \sigma(d)\}$, satisfies

$$\hat{\Omega} \equiv \arg\max_{\Omega} \sum_{i=1}^{N} \ln\left(\sum_k p_k \ell(\Omega; \boldsymbol{M}_i, w_i, z_i, d_i, k)\right)$$

where $\ell(\Omega; \boldsymbol{M}_i, w_i, z_i, d_i, k) = \pi(z,d\,|\,k) f_m(\boldsymbol{M}\,|\,k) f_w(w\,|\,k)$.

The sum inside the logarithm prohibits sequential estimation of the parameters in Ω.

Arcidiacono and Jones (2003) show the same $\hat{\Omega}$ satisfies

$$\hat{\Omega} \equiv \arg\max_{\Omega} \sum_{i=1}^{N} \sum_{k=1}^{K} p_i(k|\Omega) \ln \ell(\Omega; \boldsymbol{M}_i, w_i, z_i, d_i, k) \tag{2.4}$$

where

$$p_i(k|\Omega) \equiv \Pr(k|\boldsymbol{M}_i, w_i, z_i, d_i; \hat{\Omega}, \hat{p}) = \frac{\hat{p}_k \ell_i(\hat{\Omega}; \boldsymbol{M}_i, w_i, z_i, d_i, k)}{\sum_{k=1}^{K} \hat{p}_k \ell_i(\hat{\Omega}; \boldsymbol{M}_i, w_i, z_i, d_i, k)}$$

and

$$\hat{p}_k = \frac{1}{N}\sum_{i=1}^{N} p_i(k|\hat{\Omega}).$$

Crucially, the right-hand side of (2.4) *lends itself to sequential estimation.*

2.3.2 Bootstrap

As in Cassagneau-Francis et al. (2021), we use a bootstrap procedure to obtain standard errors and confidence intervals for our estimates to account for the random nature of the estimation algorithm. We follow the advice of O'Hagan et al. (2019) who recommend using a weighted-likelihood bootstrap (WLBS) to prevent groups from disappearing in any

[18] We could allow for heteroscedasticity here, i.e. for the variance of wages to depend on type as well as education, but we found that the algorithm performs better when only the mean of wages varies across types, and not the variance. This may be a consequence of the relatively small samples that we use in the application.

samples. The WLBS involves drawing N positive, non-zero weights from the Dirichlet distribution (which sum to N) ensuring that no observations are completely dropped from any sample. The procedure is computationally intensive as it involves re-estimating our model on 500 such weighted samples. To speed up the procedure and ensure consistent labelling we use the full-sample model estimates as starting values for each bootstrap estimation. We obtain 500 "bootstrapped estimates" for each of our model parameters and can obtain standard errors as standard deviations of these bootstrapped estimates, and confidence intervals from the corresponding quantiles.

2.4 Application: returns to a UK university degree

We now turn to our application, in which we estimate the returns to a university degree in the UK, as a function of prior ability. To achieve this aim, we apply the framework described in section 3.2 to data from the 1970 British Cohort Study (BCS 1970), following the estimation strategy outlined in section 2.3. We first briefly describe the context of higher education in the UK at the time our data was collected, then discuss the data and the specific variables used to estimate the model. The results follow in section 2.5.

2.4.1 Higher education in the UK

The higher education system in the UK has a number of features that make it well suited to studying the returns to *a university degree in general*, as we do in this paper, rather than taking a more granular approach allowing for different types of institutions and degrees. The institutions in the UK were relatively homogenous in what they offer to students. All degree-granting institutions are privately run, and in receipt of government funding.[19] The standard degree offered by a UK university is a three-year bachelor's degree specialising in a single subject, with students generally entering university at age 18 or 19.[20] The student is then awarded a Bachelor of Arts (BA, in arts or humanities subjects) or a Bachelor of Sciences (BSc) in that subject upon graduation.[21] Most universities offer students a wide range of subjects, and have large student bodies, with the largest having nearly 19,000 undergraduate students enrolled in 1994 (HESA, 1996).

A crucial difference across institutions is in their selectivity; universities were able to select students based on their prior attainment at school (as well as at interviews). Therefore, any differences across universities are not likely to be separable from differences in prior ability, a dimension which we explicitly allow returns to vary across in our

[19]There was one university in the UK which did not receive government funding at this time, and was instead run as a charity; the University of Buckingham.

[20]Figure E1 in appendix 2.E contains a detailed timeline of the university application process.

[21]There are a number of other subjects that have their own official title and abbreviation, such as the Bachelor of Laws (LLB), and Bachelor of Engineering (BEng).

analysis. Allowing for different types of institution by selectivity would likely not change our analysis. A consequence of the stratifying of universities by ability, however is that we cannot rule out that differences across individuals with different abilities are due to their attendance at different *institutions* and not due to *interaction between ability and higher education*. Separating these effects remains a question for future research.

There were no tuition fees for domestic students during the period when young people in our sample were at university, and there was a system of means-tested grants and loans for "maintenance", i.e. designed to help students cover their living costs (Greenaway and Haynes, 2003). In addition the dropout rate is particularly low, with around 90% of students completing the degree they started in 1989/1990 (Smith and Naylor, 2001). In the UK, leaving home to attend university is a major part of the experience. In the late 1980s and early 1990s when our cohort members were most likely to be at university, over 90% of university students did not live at home (HEFCE, 2009).

In terms of demographics, 46% of women and 49% of men at the "typical age of graduation" in 1996 held an undergraduate degree, with over 13% of both genders also holding a higher degree (OECD, 1998, p. 200). Splitting the population by social class, in 1991 less than 10% of those whose main parent was in the three lowest social classes[22] enrolled in university, while over 35% of children of parents in intermediate roles, and over 50% of children of professional parents enrolled in university (Dearing, 1997). In this paper we abstract from any analysis by family background, focusing only on the role of ability.

2.4.2 Data

Our data is from the 1970 British Cohort Study (BCS), an ongoing longitudinal cohort study of every person born in the United Kingdom in a week in April 1970. There were 16,568 initial cohort members (CMs), who have been contacted roughly every five years since their birth, with eleven completed "sweeps" to date. The latest sweep is currently underway in 2021 (with the CMs aged 51). In each sweep the CMs (and/or their families) are interviewed about their current circumstances and daily life, with more specific focuses at different stages of their lives. Relevant for the analysis in this paper are measures of cognitive (reading and mathematics tests) and non-cognitive (locus-of-control, self-esteem, mental health) abilities from age 16, and information on qualifications and wages at age 26.

We therefore focus on the fourth sweep,[23] which took place in 1986 (when the CMs were

[22]In the UK the six social classes are: professional (I), intermediate (II), skilled non-manual (IIIn), skilled manual (IIIm), partly skilled (IV) and unskilled (V).

[23]The fourth sweep was called *Youthscan* at the time, and the data collection was carried out by the *International Centre for Child Studies*. Information was collected from the cohort members themselves, their parents, and their schools (teachers and head teachers). The survey instruments used include questionnaires (both face-to-face and self-completion), medical examinations, diaries, and educational

Table 2.1: Description of variables used to estimate our model

Variable	Description
Wage, W	Usual weekly wage in GBP reported by the cohort member if employed at age 26. $w \equiv \log(W)$.
Cognitive score, M^C	Mean standardised score (out of 100) across reading and mathematics tests taken by the cohort members as part of the study at age 16.
Non-cognitive score, M^{NC}	Mean standardised score (zero mean, unit variance) across three measures of "personality": self esteem, locus of control, and the general health questionnaire.†
Desire to leave home, z	Response of cohort member at 16 to the question: "How much do you think [living away from home] will matter to you as an adult?".‡
Education, d	A indicator for whether the cohort member reports holding at least an undergraduate degree at age 26.

Notes: Conti and Heckman (2010) use similar measures from the same dataset to capture non-cognitive abilities and their effects on later health outcomes. †The general health questionnaire (GHQ) is a series of questions designed to predict susceptability to mental health issues. ‡Possible responses: "Matters very much"; "Matters somewhat"; "Doesn't matter".

aged 16), and on the fifth sweep which took place a decade later in 1996 (when the CMs were aged 26). These sweeps provide information from just before the decision to attend university, which is generally made at age 17 in the UK, and from when the majority of young people who would attend university have completed their degrees and entered the labour market. We split the sample by gender and estimate the model separately for men and women to enable comparison with previous work on the returns to education during this period, and because there is a significant gender pay gap in the data, both for graduates and non-graduates. Investigating the mechanisms behind the gender pay gap, though interesting and vital, is beyond the scope of this paper.

Table 2.1 describes the variables that we use to estimate our model. To arrive at our subsample, any CMs who did not respond at either the age 16 or age 26 sweep are dropped. Cohort members are also dropped if they: were not working at age 26; did not take a reading nor a maths test at age 16; did not provide responses to *any* of the non-cognitive measures;[24] are missing information on their highest qualification; or did not respond to the question about leaving home. Finally we trim the sample on wages to keep only those observations with wages between the 1st and 99th percentiles. Table 2.2 contains summary statistics for the subsample we use for our analysis, pooled and split by education and gender. Wages (W) are increasing in education (d) for both men and women. We denote log-wages by w. The mean graduate ($d=1$) wage for women is

assessments.

[24]We keep individuals for whom we have an incomplete set of cognitive or non-cognitive scores and compute the mean of non-missing scores.

Table 2.2: Summary statistics for the analysis subsample split by sex and education

Gender:	All	Male			Female		
Education:		All	$D=0$	$D=1$	All	$D=0$	$D=1$
Weekly wage (age 25, GBP, W)							
Mean	239	286	263	334	209	197	241
Std dev.	408	480	481	474	350	388	212
Degree	0.29	0.32	0	1	0.27	0	1
Male	0.40	1	1	1	0	0	0
Ability measures (M)							
Cognitive	57.0	57.8	54.3	65.3	56.5	53.6	64.4
Reading	46.1	46.0	43.0	52.4	46.1	43.6	53.1
Mathematics	40.8	41.4	38.6	48.2	40.5	38.1	47.2
Noncognitive	0.13	0.09	0.01	0.26	0.16	0.10	0.34
Self-esteem	16.5	16.5	16.1	17.3	16.5	16.2	17.1
Locus-of-control	14.3	14.4	14.1	15.0	14.3	14.2	14.5
GHQ†	1.55	1.26	1.23	1.33	1.74	1.60	2.13
Leaving home matters... (z)							
...very much	0.19	0.14	0.13	0.17	0.22	0.20	0.27
...somewhat	0.48	0.47	0.46	0.50	0.48	0.47	0.50
...doesn't matter	0.33	0.39	0.41	0.34	0.30	0.32	0.23
N	1876	745	509	236	1130	827	304

Notes: The values in the table are the mean value of that variable among the population indicated by the column headings, unless otherwise specified. The notation used in the model is in parentheses on the table to highlight which variables in the data correpond to which in the model.

below that of non-graduate men, despite women having similar cognitive test scores and higher non-cognitive measures, supporting our decision to estimate the model separately for men and women. Cognitive (M^C) and non-cognitive (M^{NC}) measures are positively correlated with education for both men and women. Our "instrument" (z), a measure of how strongly an individual wishes to leave home, is positively correlated with holding a degree ($d=1$) at age 26.

We say "instrument" as z is subject to much weaker exogeneity requirements than a usual instrument. In fact, z need not be an instrument for schooling at all. The conditions z must satisfy are that: (i) it is correlated with type (k) *or* education (d), and (ii) it is independent of wages conditional on k and d. In our application z is an instrument, and therefore we only require that z is independent of wages conditional on type (and education).

Table E1 in the appendix presents the results of a balancing exercise to provide evidence on the validity of the instrument. This exercise consists of a series of regressions with key (excluded) characteristics as the dependent variable in each regression, and the cognitive and non-cognitive measures, an indicator for females, and our instrument as co-

variates. The dependent variables are: parental income (in bands); father's (or mother's if father is absent) social class; self-assessed health; whether the young person lived in a city, town, village, or the countryside; and whether the young person is white. These are all observed at age 16. The results are reassuring, with the majority of the coefficients on the instrument not statistically significant, even at the 10% level. Self-assessed health at age 16 is the only exception. Table E2 (also in the appendix) contains the results of a multinomial logit with the young person's region of residence as the dependent, and suggests no evidence of correlation between region and the instrument once we control for prior ability using our measurements. Our balancing exercise suggests the desire to leave home is uncorrelated with other characteristics that might determine wages conditional on prior ability, and is a valid "instrument" for our purposes.

2.5 Results

This section presents the results of estimating our model on data from the BCS 1970 as we described in the previous sections.[25] We first discuss how to choose the number of types, K, which is also the number of points of support for the distribution of prior ability. We then present results by type using only cognitive ability measurements. This is not our preferred specification. However, non-cognitive measurements are rare, especially in adminstrative datasets, and therefore it is informative to see how our method performs with only cognitive measures. We then estimate our preferred specification which includes measures for both cognitive and non-cognitive prior ability, and finally we compare aggregate results across specifications, and to estimates obtained using more standard estimators.

Throughout this section we label types so that k is increasing in the mean wages of those without a degree, $\mu(k, 0)$. By estimating the model separately for men and women, we may estimate a different set of types for men and women, especially if prior ability and wage distributions differ across genders. However, we can compare types within and across genders using the type-conditional means of ability, $\alpha_C(k)$, and of wages, $\mu(k, d)$.[26]

[25] We use the sequential EM algorithm presented in section 2.3 and appendix 2.D to maximise the sample likelihood (2.4). We run *kmeans* on $\boldsymbol{M}$ and w to obtain starting values for $\pi(k)$, $\alpha(k)$, and $\mu(k, d)$. We also tried using different starting values and selecting the results with the highest likelihood, but using *kmeans* always produced a likelihood at least as high as the best among the randomly chosen starting values. We use the R programming language to implement our method. The algorithm is relatively fast to converge in our application, taking under one minute on a laptop with a quad-core Intel Core i7-6560U CPU (2.20GHz) processor and 16GB of memory, running Linux (Fedora OS). The variables we use as w, $\boldsymbol{M}$, z, and d are detailed in table 2.1.

[26] Recall that under our model of (noisy) measures and outcomes, the type-conditional means are directly informative of prior ability, although individual measures and outcomes are not.

2.5.1 Choosing K

The econometrician must set a number of types, K, before estimating the model, and so we estimate the model for K between 2 and 20 and use a range of criteria to select the best choice. What we call the *likelihood criteria*, displayed in figure 2.1a for the model with only cognitive ability measures estimated on men, are the log-likelihood and the penalised log-likelihood. We are looking for elbows where the slope of the plotted line decreases (all criteria) or maxima (AIC, BIC). The plots in figure 2.1a suggest picking a value of K less than 7, although the BIC is uninformative. We also study the aggregate results for different K to see if there are any clear patterns, or whether any values of K appear to produce anomalous results. Figure F7 in the appendix is an example of how estimated aggregate returns vary with K.

We also use as a criterion the entropy of the assignment to groups: the uncertainty or "fuzziness" in the assignment, which we assess by studying the distribution of the posterior probabilities, $p_i(k)$. Stronger assignments have clearer modes of the posterior probability distribution at 0 and 1. Figure 2.1b displays the distribution of posterior probabilities for the same model as figure 2.1a estimated on women. We would choose $K = 2$ or 3 based on this evidence. The likelihood criteria and posterior probability distributions for other models and samples are shown in appendix 2.F.1. In general the likelihood and entropy criteria suggest we want to select the lowest values of K which capture key patterns in the results.

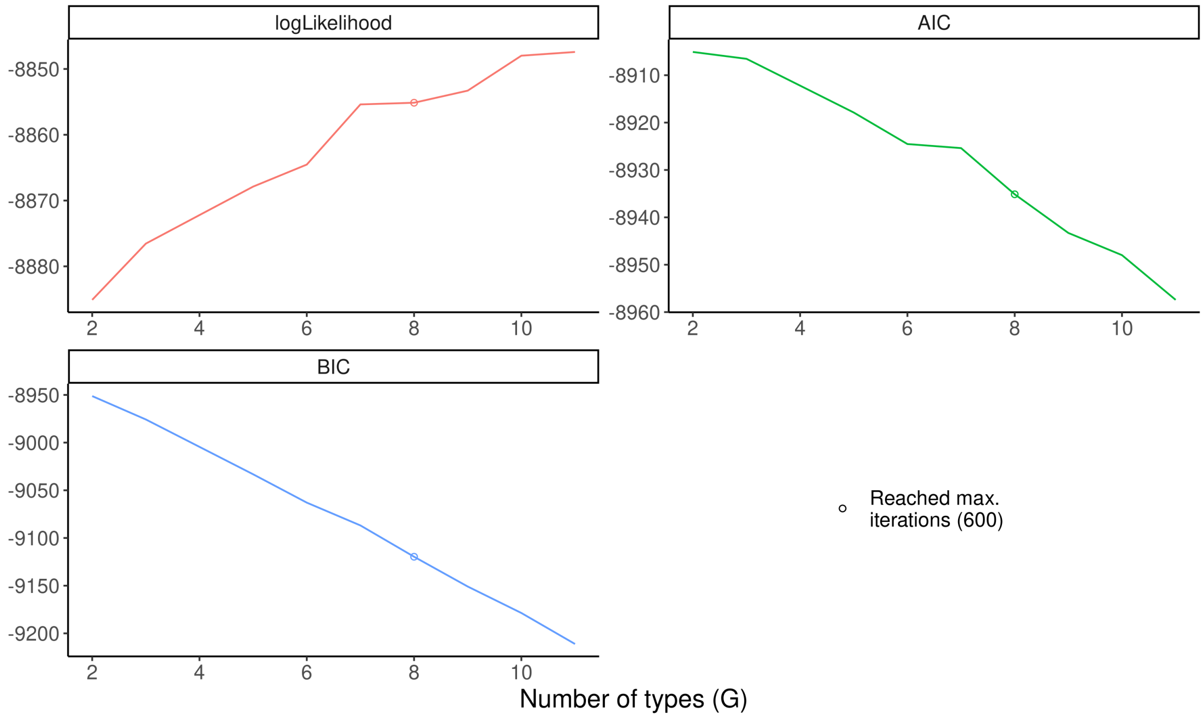
2.5.2 Measures of cognitive ability only

We first present estimates obtained using only cognitive measures of ability. Although this is not our preferred specification, datasets containing only measures of cognitive ability are generally much more widespread than those containing non-cognitive measures (or both), especially in large administrative datasets (so-called "big data"). Therefore, comparing our method using only cognitive measures with our preferred specification is important for understanding the possible limitations of these much larger (in terms of observations), though much less rich (in terms of information) datasets.

Table 2.3 displays these results for $K = 3$. The results for other values of K, which are broadly similar, are in the appendix (figure F5). Although there is a significant gender wage gap, each of the three types are close in terms of cognitive ability between males and females. For example type 1 men have a mean cognitive score of approximately 45, while type 1 women have a mean cognitive score of 40. Despite having lower wages and similar cognitive scores, the mean non-cognitive scores of the female types are higher than those of the equivalent male type. This highlights the importance of studying men and women separately as they likely face different prices for their abilities on the labour market. Table 2.3 also splits the type-conditional means by education, d. For the cognitive measure

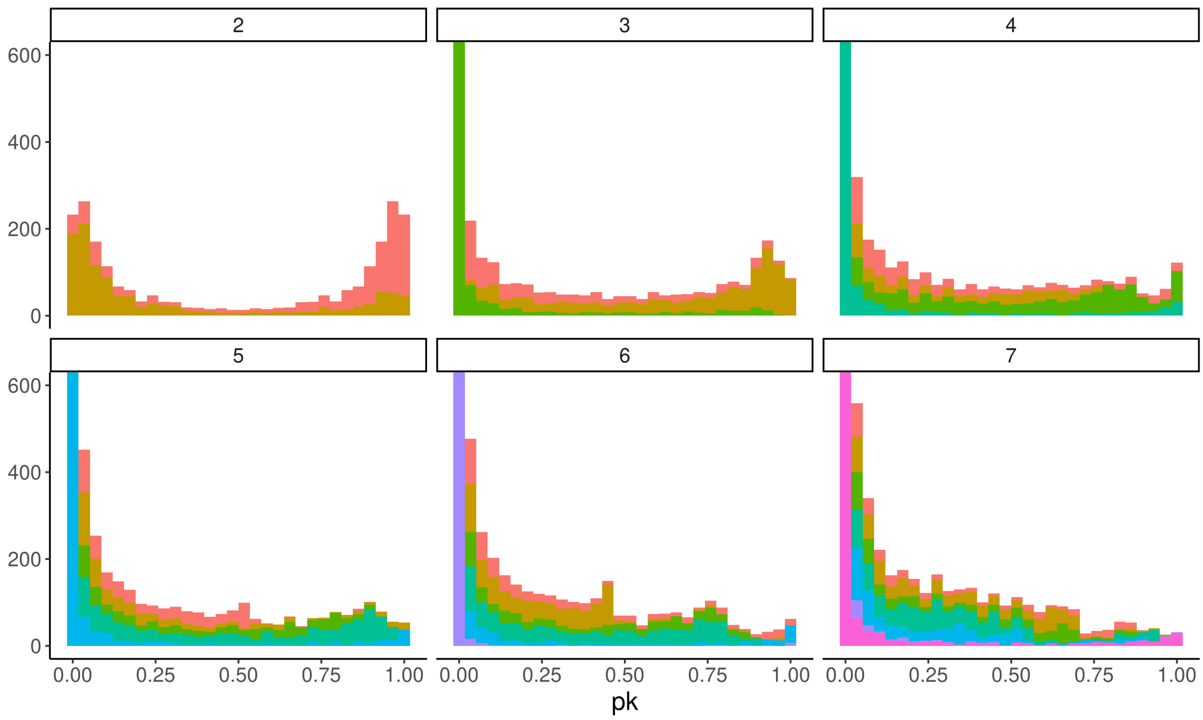
Figure 2.1: Criteria to select the number of types, K

(a) Likelihood criteria (cognitive measure, males)



Notes: In the top-left panel ("logLikelihood") we plot the loglikelihood (L) of the model against the number of types. In the top-right panel ("AIC") is the negative of the Akaike Information Criterion (AIC), with $AIC = \ln L - 2k$, where k in the number of free parameters. Finally the bottom-left panel ("BIC") plots the negative of the Bayesian Information Criterion (BIC), with $BIC = \ln L - \frac{k}{2}\ln(n)$, with n the number of observations. We are looking for "elbows" (all) and maxima (AIC and BIC). The hollow circles indicate instances in which the algorithm had not converged within 400 iterations.

(b) Distribution of posterior probabilities (cognitive measure, females)



Notes: Each panel represents a different number of types, K. The bars show the distributions of posterior probabilities, $p_i(k)$, coloured according to the value of k. Up to $K = 4$ there are modes at 0 and 1.

Table 2.3: Results by type ($K = 3$, cognitive measures only)

(a) Male

Type (k)	1		2		3	
Degree	0	1	0	1	0	1
Return to a degree	0.179		0.140		0.239	
Wage (age 25, GBP)						
Mean	205	246	221	254	230	292
Ability measures, $\mathbb{E}[M^{\ell}\vert k,d]$						
Cognitive	44.0	43.4	60.8	60.6	77.2	77.4
Non-cognitive	-0.07	0.23	0.06	0.25	0.16	0.28
$\pi(k)$	0.32	0.03	0.32	0.17	0.05	0.12

(b) Female

Type (k)	1		2		3	
Degree (d)	0	1	0	1	0	1
Return to a degree	0.222		0.286		0.188	
Wage (age 25, GBP)						
Mean	147	184	158	210	188	227
Ability measures, $\mathbb{E}[M_j\vert\theta_k,d]$						
Cognitive	40.1	36.8	58.3	58.0	77.8	77.6
Non-cognitive	0.03	0.34	0.12	0.35	0.25	0.31
$\pi(k)$	0.21	<0.01	0.49	0.17	0.02	0.09

Notes: The tables in panel (a) and (b) present the key parameter estimates from our model, and their transformations. The returns are in log-differences and are simply the within-type difference between graduate and non-graduate mean log-wages ($\mu(k,1) - \mu(k,0)$). The mean wages at 25 are the type-conditional mean log-wages exponentiated to give weekly wages in GBP, $\exp[\mu(k,d)]$. The cognitive and non-cognitive scores are simply the estimated type-conditional means, and the type proportions are the mean across all men or women of the posterior probabilities, $p_i(k)$, for each type, k.

used to estimate the model, the mean functions appear to be independent of education. However, the non-cognitive ability measure *is* correlated with education even *within* types.

Returns to a degree at each level of prior ability are generally higher for women then men, except for those with the highest cognitive ability. The pattern of returns across prior (cognitive) ability also differs across genders. The pattern for men is U-shaped, with those in the middle of the prior ability distribution experiencing lower returns than those with high or low cognitive ability. For women returns are hump-shaped with respect to prior cognitive ability, with middle types enjoying the highest returns to university. These patterns are clearest in figure F5 in the appendix. These non-linearities in returns with respect to cognitive ability highlight the importance of a framework such as ours which does not impose linearity. Given the correlation within types between non-cognitive ability and education, we will withold judgement on whether this pattern of returns to university is robust to the inclusion of both cognitive and noncognitive measures until we present the results of our preferred specification.

This correlation between non-cognitive ability measures and education is apparent in figure F6, where each type is plotted in the space of cognitive and non-cognitive skills. For both men (F6a) and women (F6a) there are large differences within types in terms of non-cognitive ability between graduates (blue) and non-graduates (red). Our method can be considered a *matching estimator*, comparing the outcomes of individuals who graduate from university, with those possessing the same latent characteristics (as captured by their ability measurements and wages) who do not graduate from (or even attend) university. A key takeaway from this analysis is that when we include only a cognitive measure in our model, we are only successful in matching along the cognitive dimension. Therefore, despite the correlation between cognitive and non-cognitive skills it appears to be important to include a measure of non-cognitive ability in the model. This raises questions about the limitations of large administrative datasets, which lack non-cognitive measures, for analyses of the returns to education. Further work is needed to determine whether there are other variables that can be used to proxy for this missing information.

2.5.3 Measures of cognitive and noncognitive ability

We now present the results of our preferred specification, which includes measures of both cognitive and non-cognitive ability. We estimated the model separately for females and males in our sample, as these groups appear to face different prices for their skills. We present estimates obtained with $K = 5$ as that is the smallest K which captures key trends and allows us to study variation in both components of ability. We discuss the results for each gender separately. However, before getting to our results we first spend some time explaining the plots in figures 2.2 and 2.3 as they are quite particular to our analysis.

The plots in panel (a) of figures 2.2 and 2.3 show the same information as those in figure

F6 for our model with cognitive and non-cognitive ability measures. The axes represent cognitive (x-axis) and non-cognitive (y-axis) ability measures, so that the location of the circles representing each type-education group reflects their relative prior ability levels. Moving north on the plot represents an increase in the mean non-cognitive ability measure, and moving east represents an increase in the mean cognitive ability measure. The sizes of the circles represent the sizes of the type-education groups, with types labelled in black text, and graduates represented by blue circles and non-graduates by red. Including non-cognitive measures was successful in one sense at least; both cognitive and noncognitive abilities are now independent of education within types. Individuals are well-matched across education groups on both cognitive *and* non-cognitive skills, in contrast to figure F6. Once again types are labelled so that k is increasing in the mean wages of non-graduates, $\mu(k,0)$.

Turning next to panel (b) of figures 2.2 and 2.3, the plots are drawn on the same axes as the plots in panel (a), so the location of the circles again reflects the mean ability measures of that type, now averaged across graduates and non-graduates of each type. However, in panel (b) the sizes of the circles represent mean *log-wages* for each group, with filled circles representing non-graduates, and hollow circles representing graduates. The difference in size between the filled and hollow circles therefore reflects the type-conditional wage return to university, a value which is also labelled on each filled circle in white. These returns are measured in log-wage differences.

Females

Our main results for women are presented in figure 2.2 and table 2.4a. Focusing first on the type-education group sizes displayed in figure 2.2a, university graduation is generally increasing in both cognitive and non-cognitive prior ability. This is reflected by the increasing size of the blue circles relative to the red as we move north-east. However, the relationship is stronger for non-cognitive prior ability. The type-conditional graduation rate,[27] denoted $\Pr(d=1|k)$ in table 2.4a, are highest for types 4 and 5, the types with the highest non-cognitive prior ability.

The presentation in figure 2.2 allows one to easily compare the relative cognitive and non-cognitive abilities for each type, and assess how varying these components affects the returns to university. However, it is difficult to see how to combine the two components into an overall measure of ability. For this we turn to table 2.4a and remind the reader that types are labelled by non-graduate wage, which is a proxy for overall prior ability, or at least a measure of the relative values of cognitive and non-cognitive abilities on the labour market. Therefore, we can study how returns vary with overall prior ability by studying how they vary across types. The general pattern of returns in table 2.4a is hump-shaped,

[27]This graduation rate is the percentage of all young people (in our analyis) of that type who have graduated from university by age 26.

with the highest returns to university being achieved by those in the middle of the prior ability distribution. This pattern is most clear in 2.4a.[28] The returns are still large at around 15 log points for those at the top and bottom of the prior ability distribution, but reach nearly 30 log points in the middle of the distribution. The graduation rates for these types is relatively low, with only 22% of those with the highest return to a degree (type 3) actually graduating from university. The graduation rate improves for type 4, the group with the next highest return, but still less than half of these young women are graduates at age 26.

Returning to figure 2.2b we can see how the different components of ability impact returns. Given the locations of the five types in cognitive-non-cognitive ability "space", it is not immediately obvious how to separate the effects of the two components (the types are not on a "grid", i.e. no two types share the same level of either component). However, some types are similar in one component. For example, moving from type 1 to type 2, involves an increase in non-cognitive ability and a slight decrease in cognitive ability. The returns to university are higher for type 2, suggesting at least at the lower end of the ability distribution increasing non-cognitive ability has a positive effect on returns.

Moving from type 2 to type 3 represents a large increase in cognitive ability and a small increase in non-cognitive ability, and results in both an increase in non-graduate wages and in the returns to a degree. Conversely, moving from type 2 to type 4 represents a small increase in cognitive ability, and a large increase in non-cognitive ability, and again we see increases in both non-graduate wages and the returns to university. This suggests that in this portion of the ability distribution cognitive and non-cognitive abilities are broad substitutes, for both graduates and non-graduates. Finally comparing types 3 and 4 to type 5, the returns to university fall, though this is driven by the relatively high non-graduate wages earned by type 5 women. These young women have the highest cognitive ability, but lower non-cognitive ability than type 4 women, suggesting high cognitive ability women have a comparative advantage as non-graduates (relative to their lower cognitive ability peers), though they still benefit from a university degree.

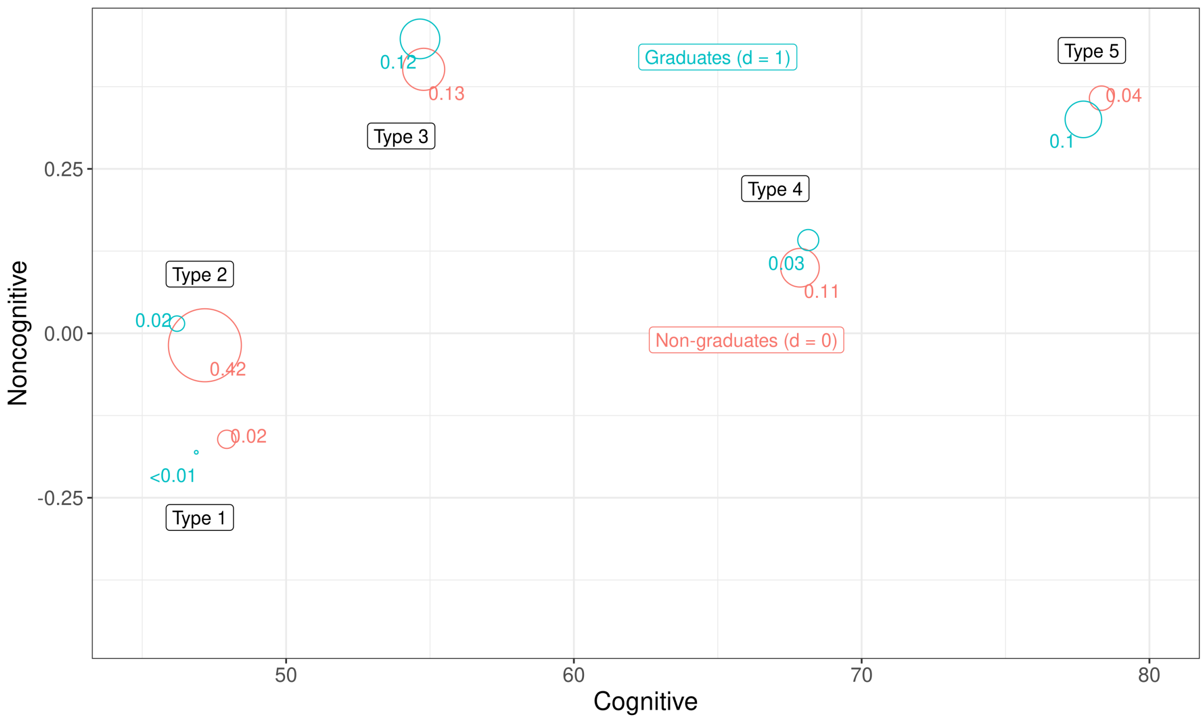
Males

We turn now to the results for young men, presented in figure 2.3 and table 2.4b. The male types follow a broadly similar pattern to those for women, with the bottom two types sharing similarly low levels of cognitive ability, and the key difference between type 1 and type 2 being their non-cognitive ability (see figure 2.3a). The university graduation rate, $\Pr(d=1|k)$ is also generally increasing with type, although type 3 men are slightly less likely to graduate from university than their type 2 peers (table 2.4b). Similar to our findings for women, non-cognitive ability seems important for gaining a university degree.

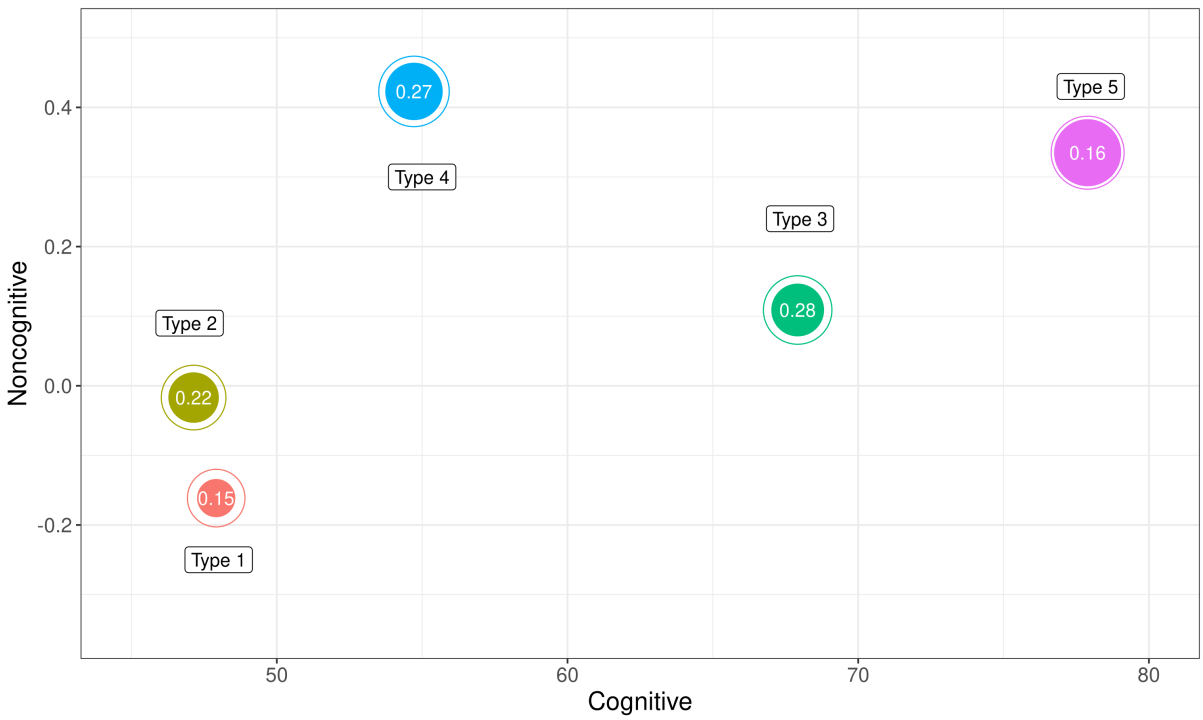
[28] The x-axis in 2.4a is not type, but type-conditional graduation rate, $\Pr(d=1|k)$. For women, this results in the same ordering as using $\mu(k,0)$.

Figure 2.2: Group sizes, locations and wage premia in cognitive-noncognitive ability space (females, $K = 5$, cognitive and non-cognitive measures)

(a) Group sizes and locations



(b) Wage premia



Notes: Panel (a) display the mean abilities (circle position) and group size (circle sizes and labels) for each type, split education (colour). Blue circles represent graduates and red non-graduates. The size of each type-education group is labelled, along with each type. Panel (b) shows the distribution of wages and wage premia by type, in the space of abilities. The positions of the circles correspond to the cognitive and non-cognitive abilities of that type. The areas of the filled circles are proportional to non-graduate log-wages, and of the hollow circles to graduate log-wages. Then the difference between the areas of filled and unfilled circles is the graduate wage premium, as a difference in log-wages. This wage premium is also labelled in white on each circle.

There is no clear pattern in returns with respect to overall prior ability, measured by non-graduate wages. If we instead order types by graduation rate, as in figure 2.4b, the returns follow a U-shaped pattern.[29] The U-shape is quite pronounced, with the least and most likely types to graduate from university enjoying returns of over 21 log points, while the "middle" types in terms of graduation rates both having returns below 15 log points. The returns are large for all types at over 11 log points, although the two types with the highest returns enjoy nearly double that of the type with the lowest return.

We can also compare types by the cognitive and non-cognitive components of their prior ability, not only their overall prior ability (non-graduate wage). Doing so, we see that types 2 and 4, whose non-cognitive ability is relatively high compared to their cognitive ability, have relatively low returns to a university degree. Moreover, those types with relatively high cognitive ability (types 3 and 5) enjoy the highest returns to university. Although prior non-cognitive skills appear to increase the likelihood of *all* young people gaining a degree, *for men* it is the interaction of prior cognitive skills with higher education that is most valued on the labour market.

Our results suggest that male graduates and non-graduates enter quite different occupations, where prior cognitive skills are better rewarded for graduates, while non-cognitive skills are (relatively) better rewarded for non-graduates. This is despite non-cognitive ability apparently increasing the likelihood of a young person graduating from university. The same cannot be said for women, whose (prior) cognitive and non-cognitive skills appear to be similarly substitutable for both graduates and non-graduates. The analysis in our paper has abstracted from considering occupations separately, including the effects of occupational choice in the "black box" that is the impact of a university degree. However, opening up this black box, including gaining a deeper understanding of how the occupations graduates choose differ from those chosen by non-graduates, will be a key focus for future research.

Marginal treatment effects. In figure 2.4 we present the type-conditional wage premiums, plotted against the type-conditional graduation rates. Presenting our results in this fashion makes them comparable to the MTE of Heckman and Vytlacil (2005). The MTE is defined by Heckman and Vytlacil (2005, p. 678) as

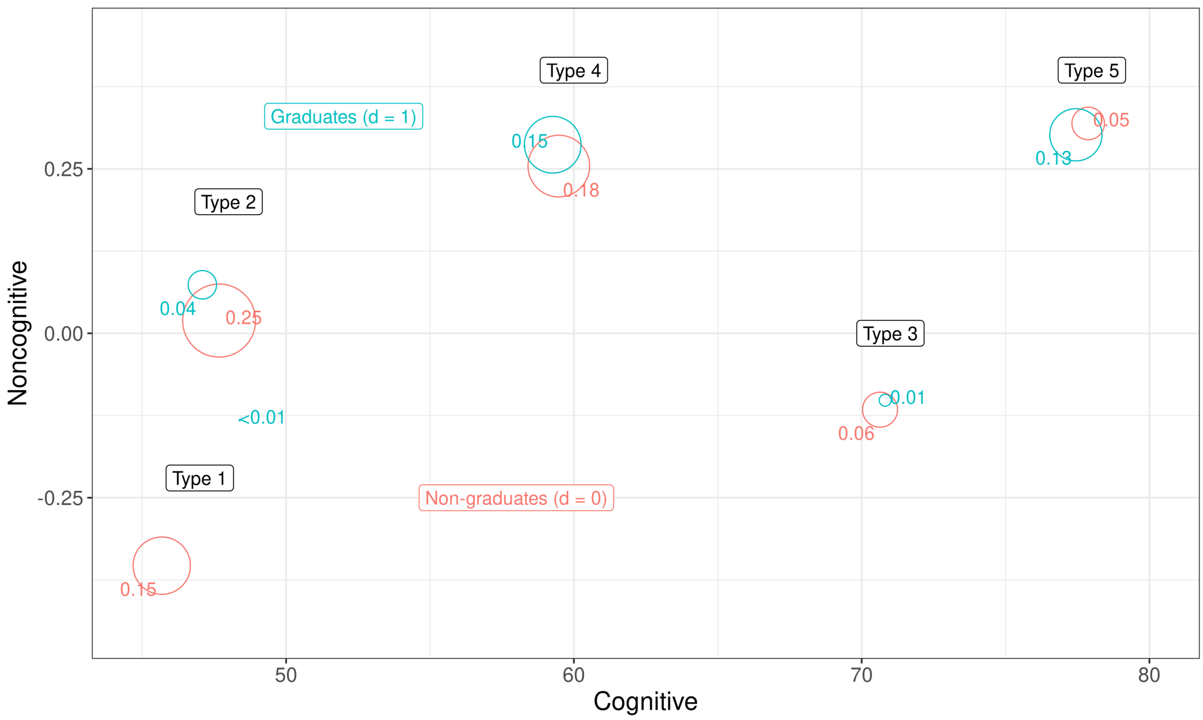
$$\Delta^{MTE}(x, u_D) \equiv \mathbb{E}[w_1 - w_0 | X = x, U_D = u_D],$$

where X are observed and and u_D are unobserved components in the decision to attend university. In our setup, a young person's type captures equivalent variation to X and u_D in Heckman and Vytlacil (2005). Therefore, our type-conditional wage premium,

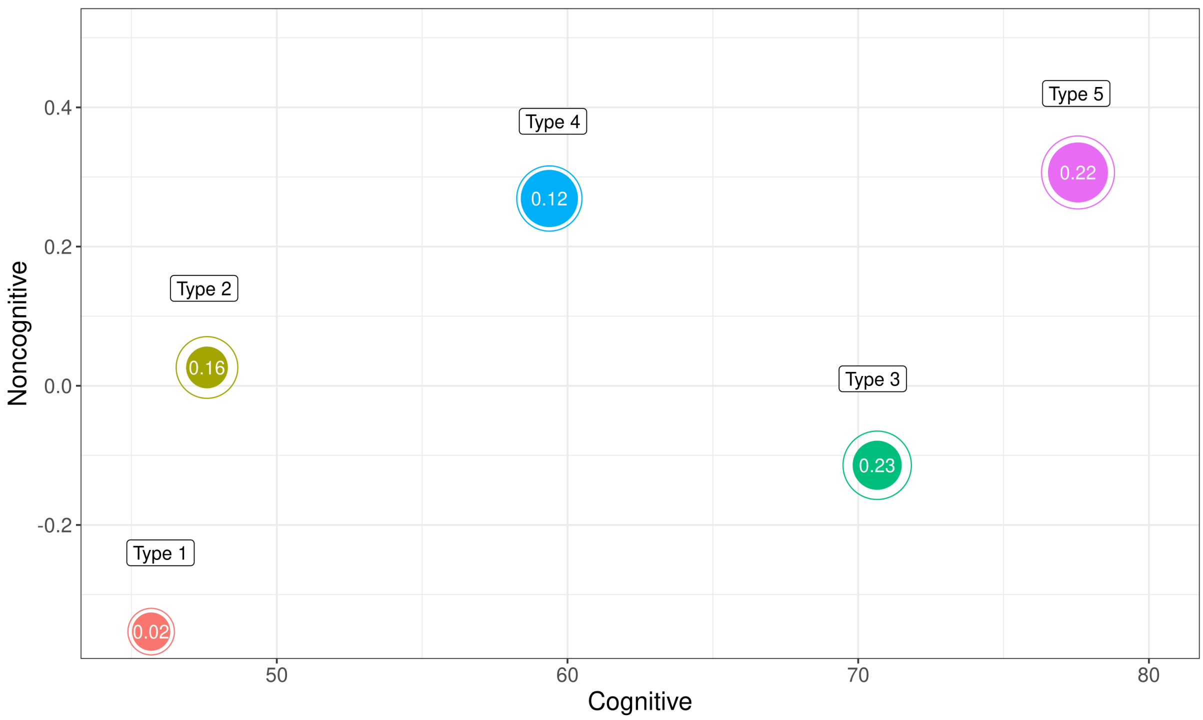
[29] This way of presenting the type-conditional returns is useful as it allows comparison with the marginal treatment effect of Heckman and Vytlacil (2005). Type 1 isomitted from this plot, in part due to its graduation rate of approximately zero.

Figure 2.3: Group sizes, locations and wage premia in cognitive-noncognitive space (males, $K = 5$, cognitive and non-cognitive measures)

(a) Group sizes and locations



(b) Wage premia



Notes: Panel (a) display the mean abilities (circle position) and group size (circle sizes and labels) for each type, split education (colour). Blue circles represent graduates and red non-graduates. The size of each type-education group is labelled, along with each type. Panel (b) shows the distribution of wages and wage premia by type, in the space of abilities. The positions of the circles correspond to the cognitive and non-cognitive abilities of that type. The areas of the filled circles are proportional to non-graduate log-wages, and of the hollow circles to graduate log-wages. Then the difference between the areas of filled and unfilled circles is the graduate wage premium, as a difference in log-wages. This wage premium is also labelled in white on each circle.

Table 2.4: Type-conditional mean ability measures and wages ($K = 5$, cognitive and noncognitive measures)

(a) Female

Type (k)	1		2		3		4		5	
Returns	0.150		0.219		0.278		0.267		0.164	
	(0.086)		(0.051)		(0.129)		(0.045)		(0.047)	
$\Pr(d=1\|k)$	0.02		0.04		0.22		0.47		0.70	
Education (d)	0	1	0	1	0	1	0	1	0	1
Wage (age 25, GBP)										
Mean	143	166	151	188	154	204	163	213	192	227
	(12.8)	(11.7)	(2.83)	(9.35)	(9.09)	(36.6)	(4.86)	(7.15)	(5.16)	(8.20)
Ability measures										
Cognitive	47.9	46.9	47.2	46.2	67.9	68.1	54.8	54.7	78.3	77.7
Non-cognitive	-0.16	-0.18	-0.02	0.01	0.10	0.14	0.40	0.45	0.36	0.33
$\pi(k,d)$	0.02	<0.01	0.42	0.02	0.11	0.03	0.14	0.12	0.04	0.10

(b) Male

Type (k)	1		2		3		4		5	
Returns	0.024		0.157		0.228		0.115		0.216	
	(0.051)		(0.108)		(0.167)		(0.067)		(0.079)	
$\Pr(d=1\|k)$	<0.01		0.13		0.09		0.46		0.73	
Education (d)	0	1	0	1	0	1	0	1	0	1
Wage (age 25, GBP)										
Mean	207	212	207	242	213	268	227	255	234	290
	(23.6)	(3.37)	(6.70)	(28.2)	(15.5)	(52.9)	(8.61)	(13.9)	(8.61)	(17.9)
Ability measures										
Cognitive	45.7	48.4	47.7	47.1	70.6	70.8	59.5	59.3	77.9	77.4
Non-cognitive	-0.35	-0.13	0.02	0.07	-0.12	-0.10	0.25	0.29	0.32	0.30
$\pi(k,d)$	0.15	<0.01	0.25	0.04	0.06	0.01	0.18	0.15	0.05	0.13

Notes: The tables in panel (a) and (b) present (transformed) key parameter estimates from our model, with bootstrapped standard errors (500 WLBS replications) in parentheses. The returns are in log-differences and are simply the within-type difference between graduate and non-graduate mean log-wages, $\mu(k,1)-\mu(k,0)$. The mean wages at 25 are the type-conditional mean log-wages exponentiated to give weekly wages in GBP, $\exp[\mu(k,d)]$. The cognitive and non-cognitive scores are simply the estimated type-conditional means, and the type proportions are the mean across all men or women of the posterior probabilities, $p_i(k)$, for each type, k.

$\mathbb{E}[w_1 - w_0|k] = ATE(k)$, is arguably analogous to the MTE. However, the MTE is usually presented ordered by u_D, not by the untreated outcome as we have done. The equivalent of ordering by u_D in our setup is to order types by $\Pr(d = 1|k)$, the type-conditional graduation rate. A strength of our framework is the flexibility in the way we model outcomes and measurements, allowing our MTE analogue to vary equally flexibly.

Evidence of non-linearity There is clear evidence of non-linearity in our results. Obtaining the combination of: (i) returns that are increasing in both cognitive and non-cognitive abilities for at least part of their distribution; while (ii) not (monotonically) increasing throughout their distribution, would not have been possible with a linear model. However, due to the correlation between cognitive and non-cognitive skills, and the apparently rather haphazard locations of the types (they do not lie nicely on a grid), determining the source of the non-linearity is difficult. It could be due to non-linearities in the returns to either component — perhaps having higher non-cognitive ability increases returns at the lower end of the distribution, while the opposite is true at the upper end — or it could be due to interactions between the components, or both. Investigating the source of these non-linearities is beyond the scope of this paper.

2.5.4 Aggregate results

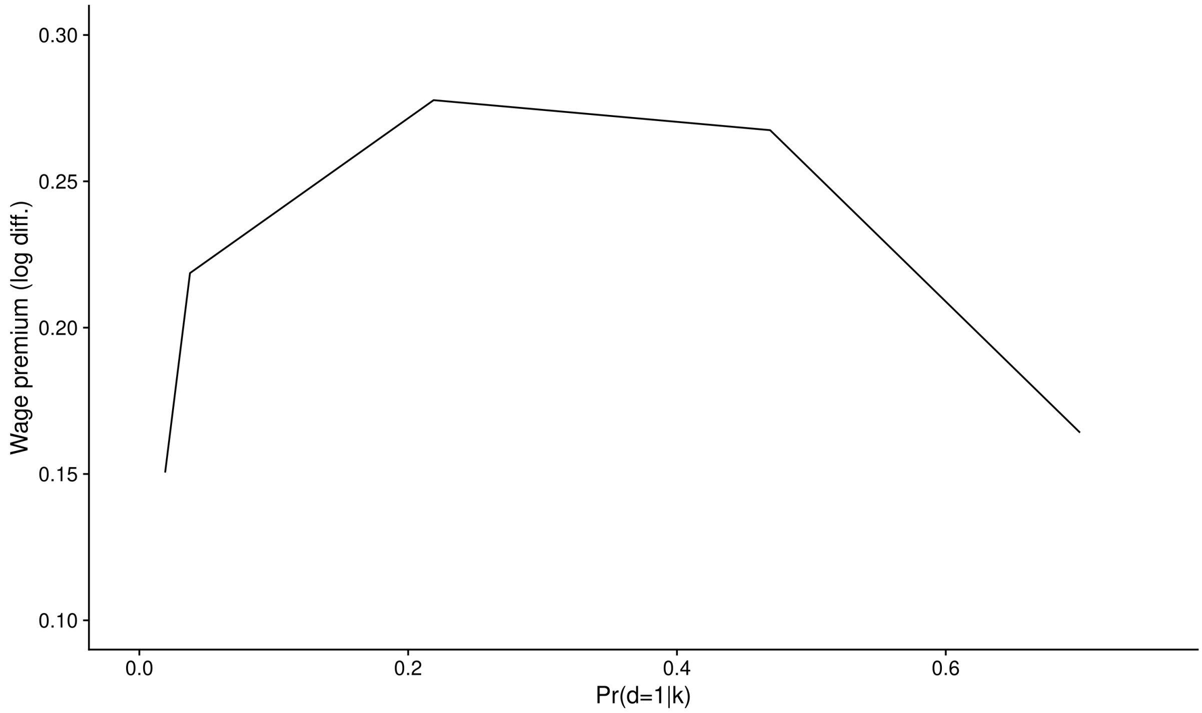
Aggregate results are not a key focus of this paper, which is primarily concerned with estimating heterogeneous returns across people with different levels of prior ability. However, it is still interesting to place our results in the context of previous work, which has generally focused on aggregate returns. In appendix 2.C we show how to aggregate across types to obtain estimates of the average returns across the whole population (ATE) and across only those who chose to attend university (ATT). These estimates are in panel (a) of table 2.5, along with standard ordinary least squares (OLS), all estimated from our model with $K = 5$. The OLS estimates calculated using our formula (b_{OLS}) are identical to those obtained from an OLS regression of log-wages on education displayed in table F5 (column 1).

Our ATE and ATT estimates are broadly similar to the OLS estimates on our data, and to estimates of other authors on UK data from a similar period.[30] Comparing our model estimates with the OLS and IV estimates using our data raises a number of points worth noting. First, the OLS estimates are broadly similar to the model estimates, though they are slightly biased relative to ATT estimates. In appendix 2.C we derive a formula for the standard OLS estimator of the return to a degree (without controls), b_{OLS}, showing how the OLS estimator is the ATT plus a bias term, B_{OLS}. We reproduce the formula

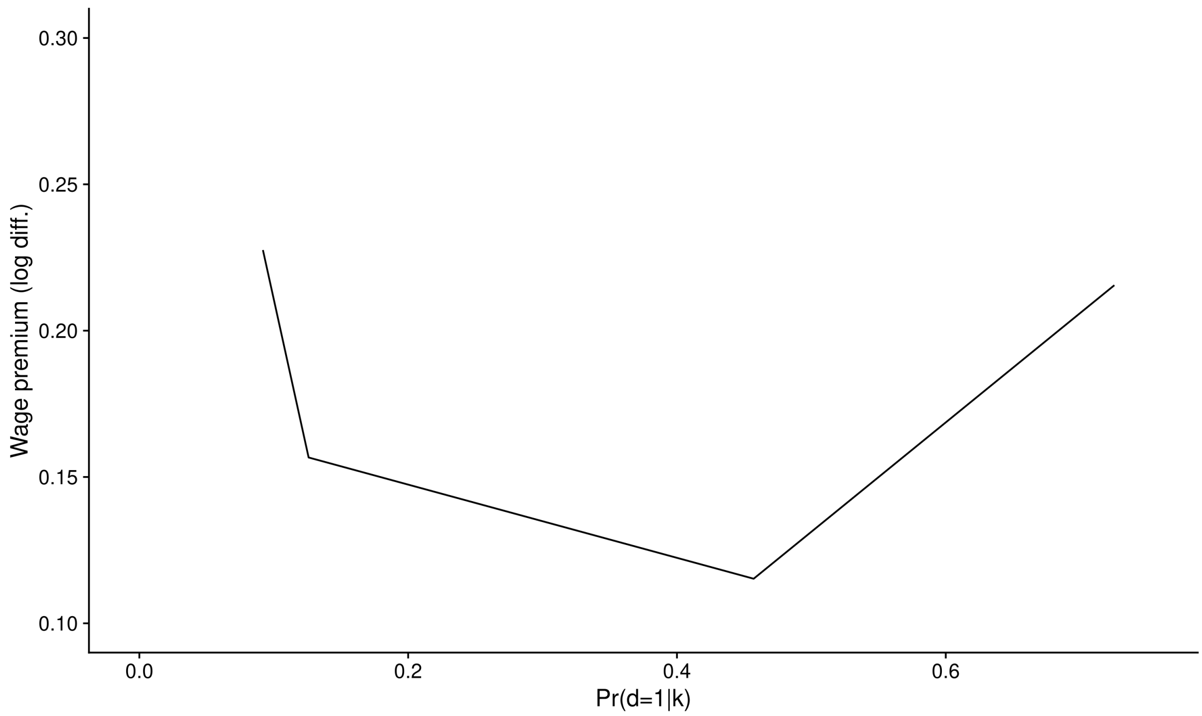
[30]Blundell et al. (2000) find wage returns of 17% for men and 37% for women at age 33 using data on a cohort born in 1958.

Figure 2.4: Returns by graduation rate

(a) Females



(b) Males



Notes: Panels (a) and (b) plot the type-conditional returns to a degree, $\mathbb{E}[w_1 - w_0|k]$ against the type-conditional graduation rates, $\Pr(d = 1|k)$. The type-conditional return for type-1 males is omitted from panel (b).

Table 2.5: Aggregate results and comparison with standard estimators ($K = 5$)

(a) Aggregate estimates

Male				Female			
ATE	ATT	b_{OLS}	B_{OLS}	ATE	ATT	b_{OLS}	B_{OLS}
0.137	0.162	0.220	0.058	0.230	0.227	0.325	0.098
(0.060)	(0.053)	(0.042)	(0.031)	(0.038)	(0.042)	(0.033)	(0.026)

(b) Weights in OLS bias (equation 2.5)

	Male					Female				
Type $k=$	1	2	3	4	5	1	2	3	4	5
Weights	-0.224	-0.252	-0.063	0.212	0.328	-0.030	-0.509	-0.037	0.259	0.316
	(0.197)	(0.158)	(0.183)	(0.204)	(0.210)	(0.097)	(0.096)	(0.018)	(0.090)	(0.086)

Notes: Panel (a) — The values in the table are calculated using the formulas in appendix 2.C. ATE and ATT are average treatment effects and average treatment on the treated. b_{OLS} is the OLS estimator, and our calculated value coincides exactly with the coefficient in a regression of wages on schooling. B_{OLS} is the bias on this estimate versus the "true" ATT. Panel (b) contains the weights used in the formula to calculate the OLS bias. Bootstrapped standard errors (500 WLBS samples) are in parentheses.

for the bias here.

$$B_{OLS} = \sum_k \underbrace{[\pi(k \mid d=1) - \pi(k \mid d=0)]}_{\text{weights}} \mathbb{E}[w_0 \mid k] \tag{2.5}$$

Panel (b) of table 2.5 contains the B_{OLS} weights estimated when $K = 5$. Some of these weights are not small, and the relatively small bias on both male and female OLS estimates appears to be due to chance: large positive and negative weights cancel each other out.

2.5.5 Comparing returns: prior ability versus university

Returning to the results in table 2.4, we can compare the effects of a low-ability individual graduating from university, with the effects of a (hypothetical) increase in their human capital. The low returns for type 1 of both genders mean they are not a good candidate for such an experiment. However, an interesting comparison involves the wages of a type 2 graduate with those of a type 5 (the highest "ability" as measured by wages) non-graduate. For men, type 2 are better off (in wage terms)[31] graduating from university than (hypothetically) increasing their prior ability to the level of the highest ability type (and not attending university). For women, type 2 would earn about the same in either counterfactual. This emphasises how high the wage returns are to a university degree, even for some lower ability young people.

[31] Here we are abstracting from the costs (both pecuniary and non-pecuniary) of graduating from university. These costs, especially the non-pecuniary or "psychic" costs, are likely decreasing in human capital and may be prohibitively high for some low ability young people.

Table 2.6: Decomposing the variance of log-wages

	Male		Female	
	Var	%	Var	%
Within (θ)	0.009	3.3	0.014	6.1
Between (d)	0.024	9.2	0.053	23.1
Unexplained	0.229	87.5	0.162	70.7
Total	0.262	100	0.229	100

Variance decomposition. The final exercise we perform with the aid of our model is to decompose the variance of wages into three parts:

"within" education groups, due to differences in prior ability;

"between" education groups, due to differences in education;

"unexplained" due to differences in individuals other than education and ability.

Formally, the decomposition is

$$\mathrm{Var}(w) = \underbrace{\mathbb{E}[\mathrm{Var}(\mathbb{E}[w\,|\,\theta,d]\,|\,d)]}_{\text{“within”}} + \underbrace{\mathrm{Var}(\mathbb{E}[w\,|\,d])}_{\text{“between”}} + \underbrace{\mathbb{E}[\mathrm{Var}(w\,|\,\theta,d)]}_{\text{“unexplained”}} \tag{2.6}$$

which allows us to compare the contributions of prior ability and of the returns to university to wage inequality. The results are in table 2.6. The majority of the variance in wages is not explained by our model. However, the contribution of the graduate wage premium ("between") to wage inequality is much larger than that of prior ability for both men and women. For women, it is particularly striking, explaining over 23% of the total variance in wages. These findings reinforce the analysis at the end of section 2.5.3 showing the wage gain from graduating for a low-ability young person (type 2) are equivalent to (hypothetically) being a non-graduate of the highest ability (type 5).

2.6 Conclusion

In this paper we have presented a framework designed to separately estimate the effects of ability and higher education on wages. We incorporate insights from the literatures on both human capital formation and the importance of non-cognitive as well as cognitive skills. Our model therefore resembles those in the literature on human capital, but one of our key innovations is a novel nonparametric identification strategy. Although we are not

the first to show non-parametric identification, our approach requires fewer measurements of prior ability than the current leading approaches in the literature. We are also the first to take an important next step, estimating our model without imposing linearity in wages nor in measurements.

We demonstrate our method in an application on data from a longitudinal cohort study in the UK. We show that a measure of cognitive ability is not sufficient to fully capture variation in (multidimensional) ability across individuals before attending university, despite strong positive correlation between cognitive and non-cognitive abilities. When we estimate our preferred specification, which includes measures of both cognitive and non-cognitive abilities, we find important non-linearities in the effects of prior ability on wages, and on the returns to a university degree. The returns to university are also shown to be more important than the returns to prior ability: a low ability young person is better off as a low-ability graduate than they would be if instead they were to increase their ability to match their highest-ability peers. The large impact of university on wages across the ability distribution leads to another of our main results: the contribution of the graduate wage premium to inequality is three to four times larger than the contribution of ability.

The implications of our findings are somewhat unsettling. We are not the first to highlight the contribution of (non-universal) higher education to wage inequality (Autor, 2014). According to our results, sending everyone (or no-one) to university would be preferable to the current situation. Moreover, given we find that returns are generally increasing in prior ability, no higher education is preferred (in terms of inequality) to universal higher education. This is clearly not a policy that many would (or should) support. However, finding ways to mitigate the contributions of higher education to inequality while preserving its many other benefits, both to the individual and society, is vital.

There are also a number of caveats to mention regarding the work in this paper. First, the framework used in this paper (and its sibling in Cassagneau-Francis et al. 2021) is new and needs to be studied in more detail. Second, this is a static, statistical analysis and so does not allow for any equilibrium considerations. Also, in our application we only consider the short term effects of higher education. In future work we plan to expand the model to allow for earnings over a longer period. Another important task is to study how the returns to higher education have evolved over recent decades. Estimating our framework on a more recent cohort would allow such an analysis.

Bibliography

Altonji, J. G. and T. A. Dunn (1996): "The Effects of Family Characteristics on the Return to Education," *The Review of Economics and Statistics*, 78, 692–704, publisher:

The MIT Press.

Arcidiacono, P. and J. B. Jones (2003): "Finite Mixture Distributions, Sequential Likelihood and the EM Algorithm," *Econometrica*, 71, 933–946, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0262.00431.

Ashenfelter, O. and C. Rouse (1998): "Income, Schooling, and Ability: Evidence from a New Sample of Identical Twins," *The Quarterly Journal of Economics*, 113, 253–284, publisher: Oxford University Press.

Autor, D. H. (2014): "Skills, education, and the rise of earnings inequality among the "other 99 percent"," *Science*, 344, 843–851.

Barrow, L. and C. E. Rouse (2005): "Do Returns to Schooling Differ by Race and Ethnicity?" *The American Economic Review*, 95, 83–87, publisher: American Economic Association.

Becker, G. (1964): *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education*, National Bureau of Economic Research, Inc.

Blundell, R., L. Dearden, A. Goodman, and H. Reed (2000): "The Returns to Higher Education in Britain: Evidence from a British Cohort," *The Economic Journal*, 110, F82–F99.

Blundell, R., L. Dearden, and B. Sianesi (2005): "Evaluating the effect of education on earnings: models, methods and results from the National Child Development Survey," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168, 473–512.

Bonhomme, S. (2021): "Teams: Heterogeneity, Sorting, and Complementarity," *SSRN Electronic Journal.*

Bonhomme, S., K. Jochmans, and J.-M. Robin (2016a): "Estimating multivariate latent-structure models," *The Annals of Statistics*, 44, 540–563, publisher: Institute of Mathematical Statistics.

——— (2016b): "Non-parametric estimation of finite mixtures from repeated measurements," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78, 211–229.

——— (2017): "Nonparametric estimation of non-exchangeable latent-variable models," *Journal of Econometrics*, 201, 237–248.

Bonhomme, S., T. Lamadon, and E. Manresa (2019): "A Distributional Framework for Matched Employer Employee Data," *Econometrica*, 87, 699–739, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA15722.

——— (2022): "Discretizing Unobserved Heterogeneity," *Econometrica*, 90, 625–643.

Bonhomme, S. and E. Manresa (2015): "Grouped Patterns of Heterogeneity in Panel Data: Grouped Patterns of Heterogeneity," *Econometrica*, 83, 1147–1184.

Bonhomme, S. and J.-M. Robin (2009): "Consistent noisy independent component analysis," *Journal of Econometrics*, 149, 12–25.

——— (2010): "Generalized Non-Parametric Deconvolution with an Application to Earnings Dynamics," *The Review of Economic Studies*, 77, 491–533, publisher: [Oxford University Press, Review of Economic Studies, Ltd.].

Bowles, S., H. Gintis, and M. Osborne (2001): "The Determinants of Earnings: A Behavioral Approach," *Journal of Economic Literature*, 39, 1137–1176.

Britton, J., L. Dearden, and B. Waltmann (2021a): "The returns to undergraduate degrees by socio-economic group and ethnicity," IFS Report.

Britton, J., L. van der Erve, C. Belfield, L. Dearden, A. Vignoles, M. Dickson, Y. Zhu, I. Walker, L. Sibieta, and F. Buscha (2021b): "How much does degree choice matter?" Tech. rep., The IFS.

Card, D. (1999): "The Causal Effect of Education on Earnings," in *Handbook of Labor Economics*, Elsevier, vol. 3, 1801–1863.

Carneiro, P., K. T. Hansen, and J. J. Heckman (2003): "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice*," *International Economic Review*, 44, 361–422.

Carneiro, P., J. J. Heckman, and E. Vytlacil (2010): "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin," *Econometrica*, 78, 377–394.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating Marginal Returns to Education," *American Economic Review*, 101, 2754–2781.

Cassagneau-Francis, O., R. J. Gary-Bobo, J. Pernaudet, and J.-M. Robin (2021): "A Nonparametric Finite Mixture Approach to Difference-in-Difference Estimation, with an Application to Professional Training and Wages," *Unpublished manuscript.*

Cawley, J., J. Heckman, and E. Vytlacil (2001): “Three observations on wages and measured cognitive ability,” *Labour Economics*, 8, 419–442.

Conti, G. and J. J. Heckman (2010): “Understanding the Early Origins of the Education-Health Gradient: A Framework That Can Also Be Applied to Analyze Gene-Environment Interactions,” *Perspectives on Psychological Science*, 5, 585–605.

Cunha, F. and J. Heckman (2007a): “The Technology of Skill Formation,” *American Economic Review*, 97, 31–47.

Cunha, F. and J. J. Heckman (2007b): “Identifying and Estimating the Distributions of Ex Post and Ex Ante Returns to Schooling,” *Labour Economics*, 14, 870–893.

——— (2008): “Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation,” *Journal of Human Resources*, 43, 738–782.

Cunha, F., J. J. Heckman, L. Lochner, and D. V. Masterov (2006): “Interpreting the Evidence on Life Cycle Skill Formation,” in *Handbook of the Economics of Education*, Elsevier, vol. 1, 697–812.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010): “Estimating the Technology of Cognitive and Noncognitive Skill Formation,” *Econometrica*, 78, 883–931, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA6551.

Dearing, R. (1997): “Higher Education in the learning society,” Tech. rep.

Delaney, L., C. Harmon, and M. Ryan (2013): “The role of noncognitive traits in undergraduate study behaviours,” *Economics of Education Review*, 32, 181–195.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977): “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the royal statistical society. Series B (methodological)*, 1–38.

French, E. and C. Taber (2011): “Chapter 6 - Identification of Models of the Labor Market,” in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card, Elsevier, vol. 4, 537–617.

Gary-Bobo, R. J., M. Goussé, and J.-M. Robin (2016): “Grade retention and unobserved heterogeneity,” *Quantitative Economics*, 7, 781–820, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/QE524.

Greenaway, D. and M. Haynes (2003): “Funding Higher Education in the UK: The Role of Fees and Loans,” *The Economic Journal*, 113, F150–F166, publisher: [Royal Economic Society, Wiley].

GRILICHES, Z. (1977): "Estimating the Returns to Schooling: Some Econometric Problems," *Econometrica*, 45, 1–22, publisher: [Wiley, Econometric Society].

HECKMAN, J. AND B. SINGER (1984): "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52, 271.

HECKMAN, J., J. STIXRUD, AND S. URZUA (2006): "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics*, 24, 411–482.

HECKMAN, J. J. AND E. VYTLACIL (2005): "Structural equations, treatment effects, and econometric policy evaluation," *Econometrica*, 73, 669–738.

HECKMAN, J. J. AND E. J. VYTLACIL (1999): "Local instrumental variables and latent variable models for identifying and bounding treatment effects," *Proceedings of the National Academy of Sciences*, 96, 4730–4734.

——— (2007): "Chapter 71 Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. 6, 4875–5143.

HEFCE (2009): "Patterns in higher education: living at home," .

HESA (1996): "Students in Higher Education Institutions 1994/95," .

IMBENS, G. W. AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467.

JACOB, B. A. (2002): "Where the boys aren't: non-cognitive skills, returns to school and the gender gap in higher education," *Economics of Education Review*, 21, 589–598.

KANE, T. J. AND C. E. ROUSE (1995): "Labor-Market Returns to Two- and Four-Year College," *The American Economic Review*, 85, 600–614, publisher: American Economic Association.

KOTLARSKI, I. (1967): "On characterizing the gamma and the normal distribution." *Pacific Journal of Mathematics*, 20, 69–76, publisher: Pacific Journal of Mathematics, A Non-profit Corporation.

MINCER, J. (1958): "Investment in human capital and personal income distribution," *Journal of political economy*, 66, 281–302.

——— (1974): *Schooling, experience, and earnings*, no. 2 in Human behavior and social institutions, New York: National Bureau of Economic Research; distributed by Columbia University Press.

OECD, ed. (1998): *Education at a glance OECD indicators*, Paris: OECD.

O'Hagan, A., T. B. Murphy, L. Scrucca, and I. C. Gormley (2019): "Investigation of Parameter Uncertainty in Clustering Using a Gaussian Mixture Model Via Jackknife, Bootstrap and Weighted Likelihood Bootstrap," *arXiv:1510.00551 [stat]*, arXiv: 1510.00551.

Reiersol, O. (1950): "Identifiability of a Linear Relation between Variables Which Are Subject to Error," *Econometrica*, 18, 375–389, publisher: [Wiley, Econometric Society].

Smith, J. P. and R. A. Naylor (2001): "Dropping out of university: A statistical analysis of the probability of withdrawal for UK university students," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164, 389–405.

Taber, C. R. (2001): "The Rising College Premium in the Eighties: Return to College or Return to Unobserved Ability?" *Review of Economic Studies*, 68, 665–691.

Todd, P. E. and K. I. Wolpin (2003): "On the Specification and Estimation of the Production Function for Cognitive Achievement," *The Economic Journal*, 113, F3–F33.

Todd, P. E. and W. Zhang (2020): "A dynamic model of personality, schooling, and occupational choice," *Quantitative Economics*, 11, 231–275.

2.A Linear model

A common assumption made to help identify and estimate models like the one above is that both wages and measurements are linear in their components. Then,

$$w_d = \mu_d^0 + \mu_d^C \theta^C + \mu_d^N \theta^N + \varepsilon_d \tag{2.7}$$
$$M_\ell = \gamma_\ell^0 + \gamma_\ell^C \theta^C + \gamma_\ell^N \theta^N + \varepsilon_\ell. \tag{2.8}$$

This is the approach taken by Cunha and Heckman (2007a,b), henceforth CH, for example.

Assumption (Independent errors). *The error terms, ε_d and ε_ℓ are independent of θ and each other, and have means equal to zero.*

Under this assumption we obtain the *classical measurement error model*, and OLS estimates using M to proxy θ as in the following equation,

$$w_d = \delta_d^0 + \boldsymbol{M}'\delta_d + \eta_d \tag{2.9}$$

where $\delta = (\delta^1, ..., \delta^L)$, are biased as

$$\begin{aligned}\mathbb{E}[\eta_d M_\ell] &= \mathbb{E}[(\epsilon_d - \boldsymbol{\varepsilon}'\delta_d)(\gamma_\ell^0 + \gamma_\ell^C \theta^C + \gamma_\ell^N \theta^N + \varepsilon_\ell)] \\ &= \delta_d \mathbb{E}[\varepsilon_\ell^2] \neq 0,\end{aligned}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_L)$, and the first equality is obtained by combining equations (2.7) and (2.8) to match equation (2.9) and equating the error terms. The second equality follows from assumption **??**. Therefore, we cannot recover θ nor any of the parameters in equations (2.7) and (2.8) via OLS. However, models with classical measurement error are well studied in economics and statistics. When w and M are *not* jointly normal, Reiersol (1950) shows that the parameters in equations (2.7) and (2.8) are identified, up to some normalisations. More recent work has shown how to identify the distribution of θ and the error terms using a theorem due to Kotlarski (1967).[32] Bonhomme and Robin (2009, 2010) generalise these results to allow for the non-parametric identification and estimation of such factor models.

2.B Nonparametric identification proof

We first state the necessary conditions under which our model is identified.

Assumption 1 (Measurements and wages). *Measurements, wages and z are independent conditional on type and education.*[33]

[32] See Carneiro et al. (2003) for more details.

[33] Measurements need not be independent of each other even conditional on type.

Assumption 2 (No empty cells). *$\pi(k,0,d) \neq 0$ for all d and for all k.*

Assumption 2 ensures that for at least one value of the instrument, arbitrarily set to zero, young people of all endowments of prior ability have positive probability of both attending and not attending university.

Assumption 3 (Linear independence). *$\left[f_M(\boldsymbol{M}|1) \quad \cdots \quad f_M(\boldsymbol{M}|k) \quad \cdots \quad f_M(\boldsymbol{M}|K)\right]$ and, for all d, $\left[f_w(w|1,d) \quad \cdots \quad f_w(w|k,d) \quad \cdots \quad f_w(w|K,d)\right]$ are linearly independent systems.*

Assumption 3 means we cannot identify any points of support in the distribution of human capital for which the associated conditional measurement and / or wage distribution can be formed by a linear combination of the distributions corresponding to other points of support. This is analogous to the rank condition in ordinary least squares.

Assumption 4 (First stage). *$\dfrac{\pi(k,1,d)}{\pi(k,0,d)} \neq \dfrac{\pi(k',1,d)}{\pi(k',0,d)}$ for all d, for all $k \neq k'$.*

Assumption 4 ensures that exposure to the instrument leads to different sized shifts in university attendance for individuals with different levels of prior ability. It is analogous to the rank condition in IV estimation.

Assumption 5 (Measurements independent of education). *For all types k and all measurements M_ℓ, $f_\ell(M^\ell|k,d) = f_\ell(M^\ell|k)$.*

Assumption 5 allows us to label groups consistently across education levels. There are other assumptions that we could make to achieve the same aim. However, it seems reasonable to assume that conditional on ability, our measurements of ability are independent of later education.

Assumption 6 (Discrete wages and measurements). *The distributions of wages and measurements have discrete support.*

Assumption 6 is not strictly necessary but it is a relatively innocuous assumption that greatly simplifies the exposition. We could discretise continuous distributions by projecting them onto some functional basis, i.e. $(\boldsymbol{M}, w) \mapsto p(z,d,\boldsymbol{M},w)$, and it is straightforward to adapt the proof.

Theorem (Identification). *Under assumptions 1-6 plus the conditional exclusion restriction on the instrument, $\pi(k,z,d)$, $f_M(\boldsymbol{M}|k) = \prod_\ell f_\ell(M_\ell|k)$, and $f_w(w|k,d)$ are nonparametererically identified.*

Proof. The proof contains three steps.

Step 1: Constructing matrices

The probability of observing an individual with variables $(z_i, d_i, \boldsymbol{m}_i, w_i)$ in our model writes

$$p(z_i, d_i, \boldsymbol{m}_i, w_i) = \sum_k \pi(k, z_i, d_i) f_m(\boldsymbol{m}_i|k) f_w(w_i|k). \tag{2.10}$$

Under assumption 6, $f_m(\cdot|k)$ and $f_w(\cdot|k)$ are probability mass functions (pmfs), which we place in matrices along with the observable probabilities, $p(z_i, d_i, \boldsymbol{m}_i, w_i)$, and the joint type-instrument-treatment probabilities, $\pi(k, z_i, d_i)$. The matrices (one per z, d value pair) containing the observed data shares,

$$\underset{n_{\boldsymbol{m}} \times n_w}{P(z,d)} \equiv \Big[p(z,d,\boldsymbol{m},w)\Big]_{\boldsymbol{m} \times w}$$

are indexed by measurement down their rows, and by wages across their columns. It has dimension $n_{\boldsymbol{m}} \times n_w$. The matrix of type-instrument-treatment probabilities for each k, z pair,

$$\underset{K \times K}{D(z,d)} \equiv \operatorname{diag}\Big[\pi(k,z,d)\Big]_{k \times k}$$

is a diagonal matrix with dimension K, containing the type-instrument-treatment probabilities, $\pi(k, z, d)$, on its diagonal. Finally, there are the two matrices containing the measurement and wage pmfs

$$\underset{n_{\boldsymbol{m}} \times K}{F_1} \equiv \Big[f_m(\boldsymbol{m}|k)\Big]_{\boldsymbol{m} \times k} \quad \text{and} \quad \underset{n_w \times K}{F_2} \equiv \Big[f_w(w|k,d)\Big]_{w \times k},$$

where F_1 is indexed by measurement down its rows, F_2 by wages down its rows, and both matrices by type across their columns. Then, $n_{\boldsymbol{m}}$ is the number of points of support in the (discrete) measurement distribution and the number of rows in F_1, and n_w is the number of points of support in the (discrete) wage distribution, and the number of rows in F_2. Both matrices have K columns.

For a given z, d, we can write equation (2.10) in matrix form

$$P(z,d) = F_1 D(z,d) F_2(d)^\top.$$

Step 2: Identifying F_1, $D(z,d)$, and $F_2(d)$

For a given d (which we omit to simplify the notation) the following matrices, corresponding to the different values of z,[34] share the same algebraic structure[35]

$$\begin{aligned} P(0) &= F_1 D(0) F_2^\top \\ P(1) &= F_1 D(1) F_2^\top \end{aligned}$$

as F_1 and F_2 are independent of z. Assumption 2 ensures $D(0)$ and $D(1)$ are invertible, and by assumption 3, the matrices F_1 and F_2 have full column rank. Therefore $P(0)$ has rank K and admits a singular value decomposition (SVD)

$$P(0) = U\Sigma V^\top,$$

where U and V are rank-n_m and rank-n_w unitary matrices. We can partition $U = \begin{bmatrix} U_1 & U_2 \end{bmatrix}$ and $V = \begin{bmatrix} V_1 & V_2 \end{bmatrix}$ so that

$$P(0) = U_1 \Sigma_1 V_1^\top, \tag{2.11}$$

where Σ_1 contains the K non-zero singular values of $P(0)$ on its diagonal. U_1 is $n_m \times K$, V_1 is $n_w \times K$, and Σ_1 is $K \times K$. From the components of the SVD in equation (2.11), we can construct the matrices $W_1 = \Sigma_1^{-\frac{1}{2}} U_1^\top$ and $W_2 = \Sigma_1^{-\frac{1}{2}} V_1^\top$. Then, applying W_1 and W_2 to $P(0)$ as follows, we obtain Q and Q^{-1}

$$\begin{aligned} W_1 P(0) W_2^\top &= \Sigma_1^{-\frac{1}{2}} U_1^\top U_1 \Sigma_1 V_1^\top V_1 \Sigma_1^{-\frac{1}{2}} = I_K \\ &= \underbrace{W_1 F_1}_{Q} \underbrace{D(0) F_2^\top W_2^\top}_{Q^{-1}} = Q Q^{-1} = I_K. \end{aligned}$$

We can similarly apply W_1 and W_2 to $P(1) := P(1,d)$, to obtain

$$W_1 P(1) W_2^\top = \underbrace{W_1 F_1}_{Q} D(1) \underbrace{F_2^\top W_2^\top}_{D(0)^{-1} Q^{-1}} = Q D(1) D(0)^{-1} Q^{-1}.$$

The non-zero (diagonal) entries of

$$D(1) D(0)^{-1} = \text{diag}\left[\frac{\pi(k,1,d)}{\pi(k,0,d)}\right]_K$$

are the (unique) eigenvalues of $W_1 P(1) W_2^\top$, which is derived using only the observable matrices P(1) and P(0). Q contains eigenvectors of $W_1 P(1) W_2^\top$, though these eigenvectors are only determined up to a multiplicative constant.

[34] This example is for a binary z, but the proof is easily extended to any discrete, finite z.

[35] By this we mean they can be decomposed into a trio of matrices, where the first and third matrices are identical (F_1 and $F_2^\top$) and the middle matrix is diagonal.

To pin down these eigenvectors, recall that U is unitary and hence

$$U_2^\top P(0) = U_2^\top U_1 \Sigma_1 V_1^\top = 0_{(n_m-K)\times n_w}$$

which implies

$$U_2^\top P(0) = U_2^\top F_1 D(0) F_2^\top = 0_{(n_m-K)\times n_w}. \tag{2.12}$$

By assumptions 2 and 3, $D(0)F_2^\top$ has full row rank, so equation (2.12) implies $U_2^\top F_1 = 0_{(n_m-K)\times n_w}$. Now define $\widehat{Q}$ as some matrix of eigenvectors of $W_1 P(1) W_2^\top$, such that there is a diagonal matrix Δ which satisfies $\widehat{Q} = Q\Delta = \Sigma_1^{-\frac{1}{2}} U_1^\top F_1 \Delta$, and $\Sigma_1^{\frac{1}{2}}\widehat{Q} = U_1^\top F_1 \Delta$. Therefore

$$\begin{pmatrix} \Sigma_1^{\frac{1}{2}}\widehat{Q} \\ 0_{(n_m-K)\times n_w} \end{pmatrix} = U^\top F_1 \Delta, \tag{2.13}$$

and

$$U_1 \Sigma_1^{\frac{1}{2}} \widehat{Q} = U \begin{pmatrix} \Sigma_1^{\frac{1}{2}}\widehat{Q} \\ 0_{(n_m-K)\times n_w} \end{pmatrix} = U U^\top F_1 \Delta = F_1 \Delta. \tag{2.14}$$

Then $F_1\Delta = U_1\Sigma_1^{\frac{1}{2}}\widehat{Q}$ is identified, and also we have that $F_1 = U_1 \Sigma_1^{\frac{1}{2}} \widehat{Q}\Delta^{-1}$. Noticing the rows of F_1 must sum to one (as each column is a probability distribution), we can find the non-zero (diagonal) elements of Δ using

$$(\Delta_1, ..., \Delta_K) = (1, ..., 1) U_1 \Sigma_1^{\frac{1}{2}} \widehat{Q}, \tag{2.15}$$

which identifies Δ and hence F_1.

Finally, $\Delta\widehat{Q}^{-1} = Q^{-1} = D(0)F_2^\top V_1 \Sigma_1^{\frac{1}{2}}$, and hence $Q^{-1}\Sigma^{\frac{1}{2}} = D(0)F_2^\top V_1$. V is an unitary matrix, so $P(0)V_2 = 0$, using that $V_1^\top V_2 = 0$, and $F_1 D(0)$ has rank K, implying $F_2^\top V_2 = 0_{n_w \times (n_m-K)}$. Then, following similar steps to before,

$$\begin{aligned}
Q^{-1}\Sigma^{\frac{1}{2}} V_1^\top &= \begin{pmatrix} D(0)F_2^\top V_1 & 0_{n_w\times(n_m-K)} \end{pmatrix} \begin{pmatrix} V_1^\top \\ V_2^\top \end{pmatrix} \\
&= \begin{pmatrix} D(0)F_2^\top V_1 & D(0)F_2^\top V_2 \end{pmatrix} V^\top \\
&= D(0)F_2^\top V V^\top \\
&= D(0)F_2^\top.
\end{aligned} \tag{2.16}$$

The rows of F_2 also sum to one, so $D(0)$ and hence F_2 are identified from equation (2.16), following a similar argument the one used above to identify Δ and F_1. And $D(1)$ is known now we know $D(0)$ and $D(1)D(0)^{-1}$.

Step 3: Correct labels across d.

We need to ensure that the labels on types are consistent across treatments (i.e. values of d). We use that F_1 is independent of d to ensure that each type is labelled the same across all treatments.

□

2.C Treatment effects in our framework

As in Cassagneau-Francis et al. (2021), here we show that we can identify some of the usual treatment effect (TE) estimands and their associated biases using our framework.

Average treatment effect. We can aggregate over types to obtain the ATE in (2.3).

$$ATE \equiv \mathbb{E}[w_1 - w_0] = \sum_k \pi(k) ATE(k)$$

where $\pi(k) = \sum_{z,d} \pi(k,z,d)$, the proportion of young people of type k.

Average treatment on the treated. We can also aggregate over those who attend university within each type to obtain the ATT.

$$ATT \equiv \mathbb{E}[w_1 - w_0 \,|\, d=1] = \sum_k \pi(k \,|\, d=1) ATE(k)$$

where $\pi(k \,|\, d=1) = \frac{\sum_z \pi(k,z,1)}{\sum_{k,z} \pi(k,z,1)}$, the proportion of individuals of type k among those who attend university.

Ordinary least squares (OLS). We can also calculate the OLS estimator, b_{OLS}, within our framework, and decompose this estimand into an ATT term and an "OLS bias" term, B_{OLS}.

$$\begin{aligned} b_{OLS} = \frac{\mathrm{Cov}(w,d)}{\mathrm{Var}(d)} &= \mathbb{E}[w_1 \,|\, d=1] - \mathbb{E}[w_0 \,|\, d=0] \\ &= \sum_k \pi(k \,|\, d=1)\, \mathbb{E}[w_1 \,|\, k] - \pi(k \,|\, d=0)\, \mathbb{E}[w_0 \,|\, k] \\ &= ATT + B_{OLS} \end{aligned}$$

where

$$B_{OLS} = \sum_k [\pi(k \,|\, d=1) - \pi(k \,|\, d=0)]\, \mathbb{E}[w_0 \,|\, k]$$

The OLS bias disappears only if (i) $\pi(k \,|\, d=1) = \pi(k \,|\, d=0)$ for all values of k; or (ii) $\mathbb{E}[w_0 \,|\, k] = \mathbb{E}[w_0 \,|\, k']$ for all $k \neq k'$. The first equality is unlikely to hold in our application

as those with higher prior are more likely to attend university, and hence the proportion of those with high prior is likely to be larger among graduates. This is the issue of selection on ability that was mentioned earlier. The second equality is also unlikely to hold, as young people with higher prior ability are generally more productive workers and can hence command a higher wage.

IV and LATE. Finally, we can perform a similar exercise to decompose the standard (two-stage least squares) IV estimator for a binary instrument, into a LATE term which corresponds to Imbens and Angrist (1994)'s local average treatment effect, and an "IV bias" term, B_{IV}.

The two-stage least squares estimator of the effect of university on wages (without controls) is

$$b_{IV} = \frac{\mathrm{Cov}(w,z)}{\mathrm{Cov}(d,z)} = \frac{\mathbb{E}[w|z=1]-\mathbb{E}[w|z=0]}{\mathbb{E}[d|z=1]-\mathbb{E}[d|z=0]}$$

In our framework, the denominator of b_{IV} is

$$\mathbb{E}[d|z=1]-\mathbb{E}[d|z=0] = \sum_k \left[\pi(k,d=1\,|\,z=1)-\pi(k,d=1\,|\,z=0)\right].$$

The numerator has a more interesting decomposition, such that we can write

$$b_{IV} = LATE + B_{IV}$$

where

$$LATE = \sum_k \frac{\pi(k,d=1|z=1)-\pi(k,d=1|z=0)}{\sum_k \left[\pi(k,d=1|z=1)-\pi(k,d=1|z=0)\right]} ATE(k), \tag{2.17}$$

and

$$B_{IV} = \sum_k \frac{\pi(k|z=1)-\pi(k|z=0)}{\sum_k \left[\pi(k,d=1|z=1)-\pi(k,d=1|z=0)\right]} \mathbb{E}[w_0\,|\,k],$$

with

$$\pi(k,d|z) = \frac{\pi(k,z,d)}{\sum_{k,d}\pi(k,z,d)} \quad \text{and} \quad \pi(k|z) = \sum_d \pi(k,d|z).$$

Therefore the *LATE* estimator of Imbens and Angrist (1994) is a weighted average of type-specific *ATE*s in our framework, with weights corresponding to the proportion of *compliers*[36] with that level of prior ability. Note the similarity between our decomposition of the *LATE* in equation (3.3) and Heckman and Vytlacil (1999, 2005, 2007)'s marginal treatment effect (*MTE*):[37]

$$LATE = \frac{\int_{\varphi(0)}^{\varphi(1)} \Delta^{MTE}(\nu)\,\mathrm{d}\nu}{\varphi(1)-\varphi(0)} \tag{2.18}$$

[36] Those who are induced into attending university by the instrument.

[37] This formula is adapted from the presentation in French and Taber (2011)'s excellent survey on the identification of models of the labour market.

where

$$\Delta^{MTE}(\nu) = \mathbb{E}[w_1 - w_0 \,|\, \nu_i = \nu],$$

and $\varphi(z)$ and ν are the observed and unobserved components of the non-pecuniary cost of attending university. The formula in (2.18) is a weighted average of the returns to university over those induced to attend by the instrument, though this average is over the distribution of (unspecified) unobserved costs, rather than (imperfectly observed) prior ability. Our framework is more flexible in one sense as it allows correlation between outcomes (w_d) and unobserved costs through latent types.

2.D EM algorithm details

The EM algorithm iterates back and forth over the following two steps:

E-step.

The E-step updates the posterior type probabilities, $p_i(k|\Omega)$:

$$p_i\left(k|\hat{\Omega}^{(s)}\right) \equiv \frac{\hat{p}_k^{(s)} \ell\left(\hat{\Omega}^{(s)}; \boldsymbol{M}_i, w_i, z_i, d_i, k\right)}{\sum_{k=1}^{K} \hat{p}_k^{(s)} \ell\left(\hat{\Omega}^{(s)}; \boldsymbol{M}_i, w_i, z_i, d_i, k\right)}, \tag{2.19}$$

where $\Omega = \{\pi(z,d|k), \alpha_j(k), \omega_j(k), \mu(k,d), \sigma(d)\}$, over all values such that $z \in \{0,1\}$, $d \in \{0,1\}$, $k \in \{1,...,K\}$, and $j \in \{C,N\}$.

M-step.

While in the M-step we update the components of Ω in the $(s+1)$-th iteration, using the estimates from the s-th iteration.

- Update $\alpha_j(k), \omega_j(k)$.

  1. Update $\alpha_j(k)$ as the weighted mean test score, using posterior probabilities as weights (for each type)

  $$\alpha_j(k)^{(s+1)} \equiv \frac{\sum_i p_i\left(k|\hat{\Omega}^{(s)}\right) M_{ji}}{\sum_i p_i\left(k|\hat{\Omega}^{(s)}\right)} \tag{2.20}$$

  2. Then $\omega_j(k)$ is updated as the weighted root-mean-square error, using posteriors as weights

  $$\omega_j(k)^{(s+1)} \equiv \sqrt{\frac{1}{N}\sum_{i=1}^{N} p_i\left(k|\hat{\Omega}^{(s)}\right)\left(M_{ji} - \alpha_j(k)^{(s+1)}\right)^2} \tag{2.21}$$

- Update $\mu(k,d)$, $\sigma(k,d)$.

1. Again use weighted means, with weights $p_i(k|\hat{\Omega}^{(s)})$ to update $\mu(k,d)$:

$$\mu(k,d)^{(s+1)} \equiv \frac{\sum_{i:d_i=d} p_i\left(k|\hat{\Omega}^{(s)}\right) w_i}{\sum_{i:d_i=d} p_i\left(k|\hat{\Omega}^{(s)}\right)} \tag{2.22}$$

2. And use the updated $\mu(k,d)$ to update $\sigma(d)$:

$$\sigma(d)^{(s+1)} \equiv \sqrt{\frac{\sum_k \sum_{i:d_i=d} p_i\left(k|\hat{\Omega}^{(s)}\right)\left(w_i - \mu_d^{(s+1)}(k)\right)^2}{\sum_k \sum_{i:d_i=d} p_i\left(k|\hat{\Omega}^{(s)}\right)}} \tag{2.23}$$

- Finally, we sum posterior probabilities by k, z, and d to obtain $\pi(k,z,d)$,

$$\pi(k,z,d)^{(s+1)} \equiv \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I(z,d)} p_i(k|\hat{\Omega}^{(s)}), \tag{2.24}$$

where $I(z,d) = \{i : z_i = z,\, d_i = d\}$.

Iterations stop when the algorithm converges, i.e. when the increase in likelihood between iterations is below a threshold:

$$\mathcal{L}(\Omega^{(s)}; \boldsymbol{M}, w, z, d) - \mathcal{L}(\Omega^{(s-1)}; \boldsymbol{M}, w, z, d) < \delta, \tag{2.25}$$

for some $\delta > 0$ chosen by the econometrician.

2.E Context and data

Figure E1: Timeline of educational decisions

Year 11 (age 15/16)	Apply for sixth form
End of year 11	Sit GCSEs & legally able to leave school
Year 13 (age 17/18)	Apply to universities & receive "conditional" offers
End of year 13	Sit A-levels
Summer after year 13	Receive A-levels results & confirm place at university

Table E1: Balancing checks for instrument validity

	Dependent variable:				
	Parental				
	Income	Social class	Health	Urban	White
	(1)	(2)	(3)	(4)	(5)
Female	−0.223**	−0.188*	0.245***	0.084	−0.008
	(0.097)	(0.100)	(0.094)	(0.092)	(0.008)
Cognitive	0.016***	0.024***	−0.005	0.004	0.002***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.0003)
Non-cognitive	0.236***	0.250***	−0.180**	0.047	0.004
	(0.083)	(0.088)	(0.080)	(0.078)	(0.007)
Leaving home...					
matters somewhat	−0.070	−0.047	0.236*	−0.133	0.006
	(0.128)	(0.131)	(0.123)	(0.120)	(0.011)
doesn't matter	−0.185	−0.126	0.069	−0.180	0.009
	(0.137)	(0.139)	(0.131)	(0.129)	(0.012)
Observations	1,398	1,418	1,870	1,822	1,816
R^2					0.023
Adjusted R^2					0.020
Residual Std. Error					0.170
F Statistic					8.452***

Notes: *p<0.1; **p<0.05; ***p<0.01
In columns (1–4) the dependent variables are categorical and ordered logit was used to regress the dependent variable on the covariates, using the `polr` function from the R package MASS. In column (5), as the dependent variable is binary we estimate a linear probability model using function `lm` from the R stats package.

Table E2: Balancing checks for instrument validity: regions

	Dependent variable:									
	North West	Yorkshire	East Mids	West Mids	East	London	South East	South West	Wales	Scotland
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Female	−0.179	−0.437*	−0.498*	−0.444*	−0.535**	−0.515*	−0.444*	0.017	−0.512*	−0.201
	(0.252)	(0.258)	(0.272)	(0.264)	(0.254)	(0.290)	(0.243)	(0.281)	(0.294)	(0.259)
Cognitive	0.009	−0.017**	−0.019**	−0.011	−0.002	−0.017*	0.005	−0.0002	−0.012	−0.004
	(0.008)	(0.008)	(0.009)	(0.009)	(0.008)	(0.009)	(0.008)	(0.009)	(0.010)	(0.008)
noncogScore	0.106	0.440**	0.404*	0.215	0.251	0.056	0.212	0.228	0.036	0.199
	(0.207)	(0.214)	(0.227)	(0.219)	(0.211)	(0.241)	(0.200)	(0.227)	(0.245)	(0.212)
Leaving home...										
matters somewhat	0.293	0.038	0.378	0.380	0.537	0.138	0.350	0.159	0.071	0.275
	(0.314)	(0.319)	(0.352)	(0.334)	(0.329)	(0.370)	(0.299)	(0.339)	(0.368)	(0.316)
doesn't matter	0.262	0.017	0.314	0.090	0.413	0.115	−0.091	0.026	−0.092	−0.177
	(0.328)	(0.332)	(0.366)	(0.353)	(0.343)	(0.385)	(0.319)	(0.358)	(0.388)	(0.339)
Akaike Inf. Crit.	8,754.993	8,754.993	8,754.993	8,754.993	8,754.993	8,754.993	8,754.993	8,754.993	8,754.993	8,754.993

Notes: *p<0.1; **p<0.05; ***p<0.01. We use the function `multinom` from the R package nnet to perform the multinomial logit used to estimate the coefficients in this table as region is an unordered categorical variable.

2.F Results

2.F.1 Choosing K

2.F.2 Single cognitive measure

2.F.3 Cognitive and non-cognitive measures

Table F1: Distribution parameter estimates (male, cognitive and non-cognitive measures)

	$K=5$									
Type(k) =	1		2		3		4		5	
$\alpha_C(k)$	45.7		47.6		70.7		59.4		77.6	
$\omega_C(k)$	8.99		13.6		4.78		9.06		6.60	
$\alpha_N(k)$	−0.35		0.03		−0.11		0.27		0.31	
$\omega_N(k)$	0.71		0.37		0.60		0.46		0.50	
$d=$	0	1	0	1	0	1	0	1	0	1
$\mu(k,d)$	5.33	5.35	5.33	5.49	5.36	5.59	5.43	5.54	5.46	5.67
$\sigma(d)$	0.46	0.52	0.46	0.52	0.46	0.52	0.46	0.52	0.46	0.52

Table F2: Distribution parameter estimates (female, cognitive and non-cognitive measures)

	$K=5$									
Type(k) =	1		2		3		4		5	
$\alpha_C(k)$	47.9		47.1		67.9		54.7		77.9	
$\omega_C(k)$	9.25		12.6		4.86		7.88		6.77	
$\alpha_N(k)$	−0.16		−0.02		0.11		0.42		0.33	
$\omega_N(k)$	1.21		0.52		0.54		0.56		0.51	
$d=$	0	1	0	1	0	1	0	1	0	1
$\mu(k,d)$	4.96	5.11	5.02	5.23	5.04	5.32	5.09	5.36	5.26	5.42
$\sigma(d)$	0.52	0.42	0.52	0.42	0.52	0.42	0.52	0.42	0.52	0.42

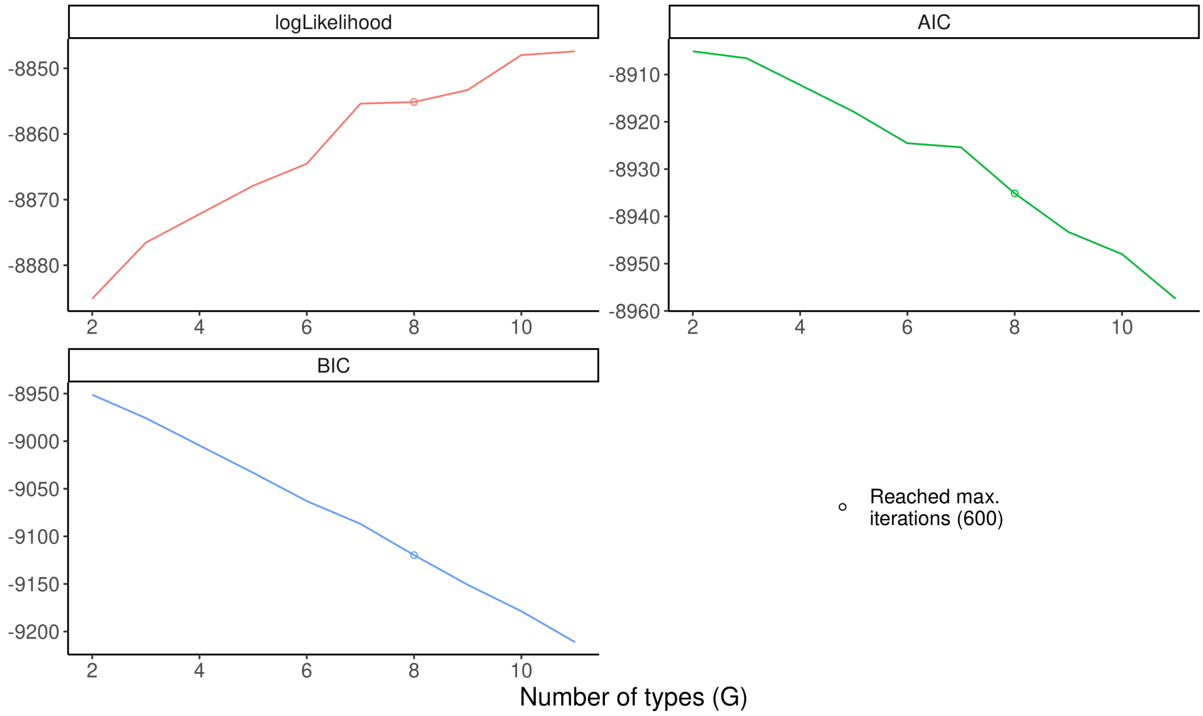
2.F.4 OLS estimates

In table F5, we present ordinary least squares (OLS, columns 1–4) and two-stage least squares (2SLS, columns 5 and 6) estimates of the returns to a university degree, across a range of specifications. The baseline regression equation is

$$w_i = \beta_0 + \mu_d d_i + \gamma_C M_i^C + \gamma_N M_i^N + X_i' \beta_1 + \varepsilon_i$$

Figure F1: Likelihood criteria: single cognitive measure

(a) Male



(b) Female

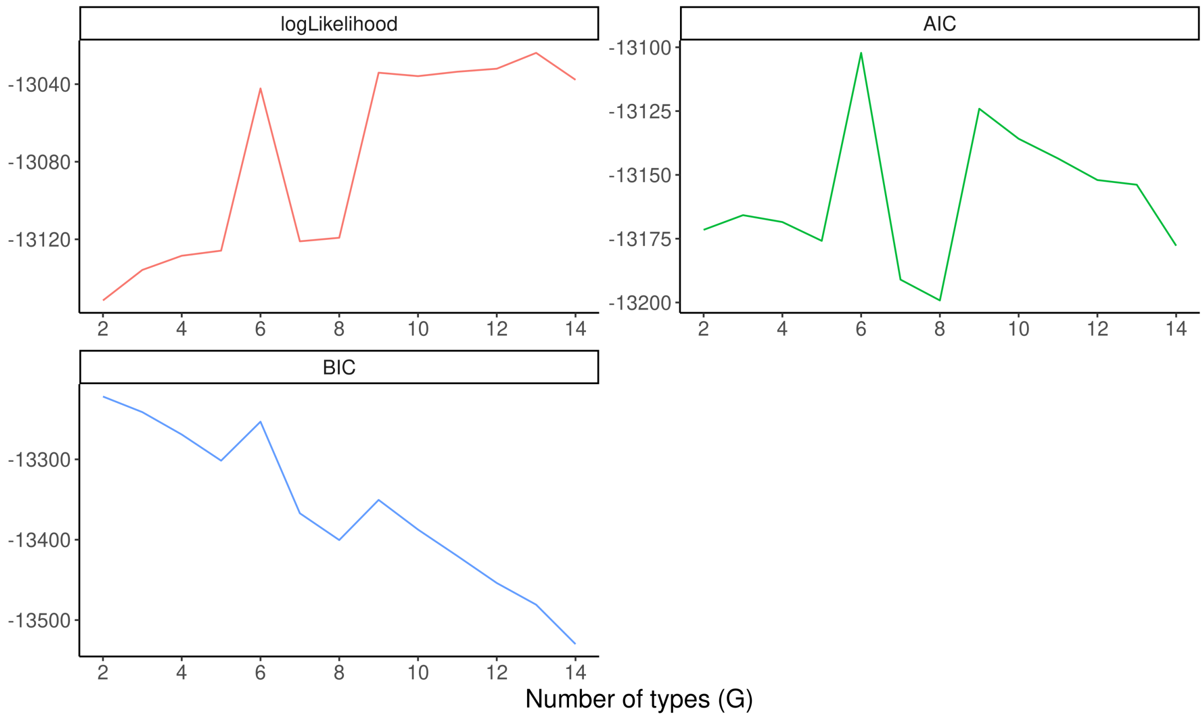
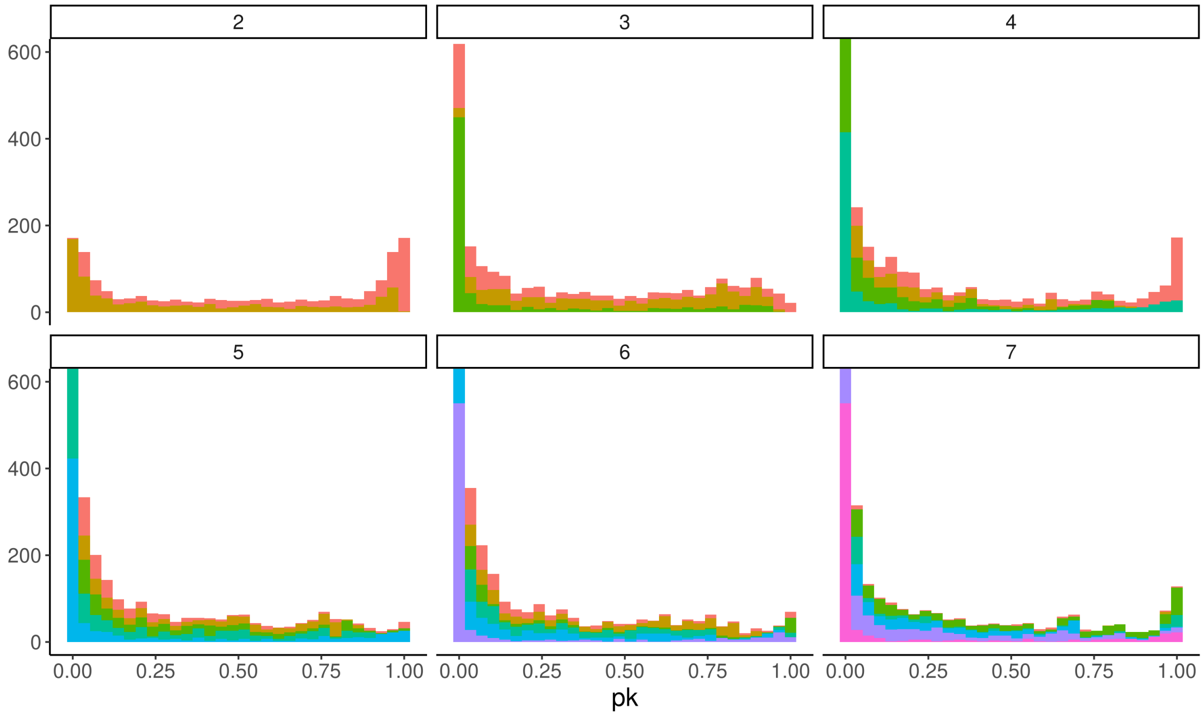


Figure F2: Posterior probabilities: single cognitive measure

(a) Male



(b) Female

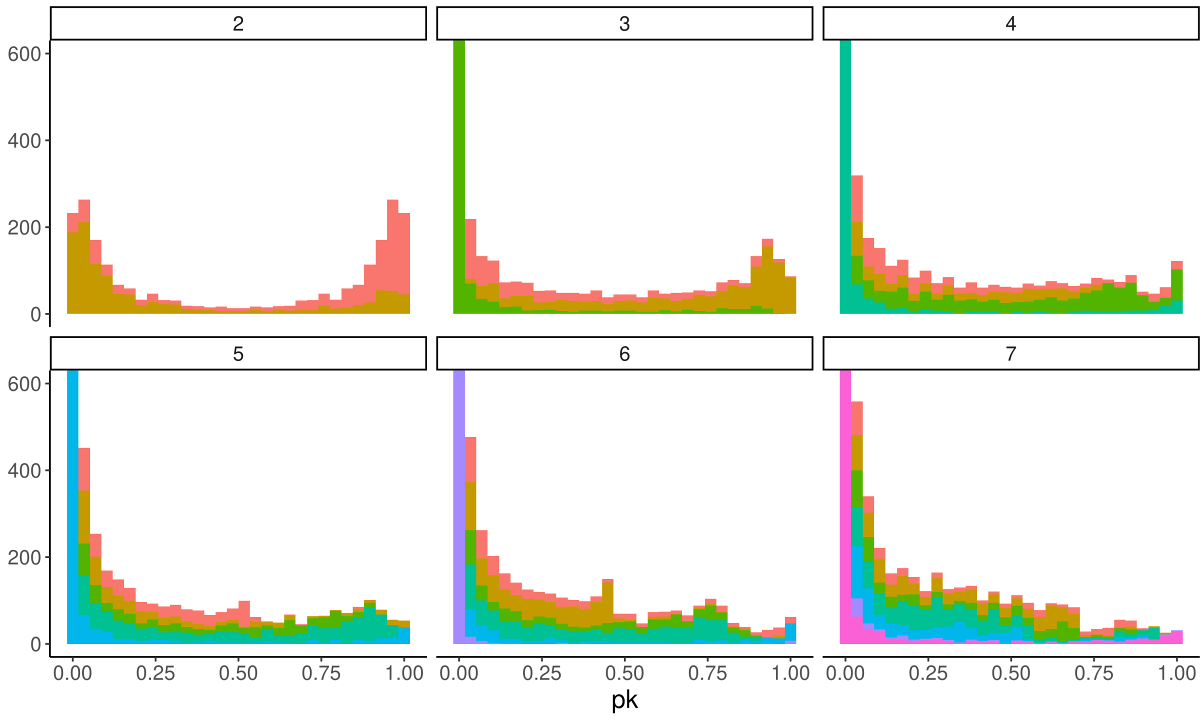
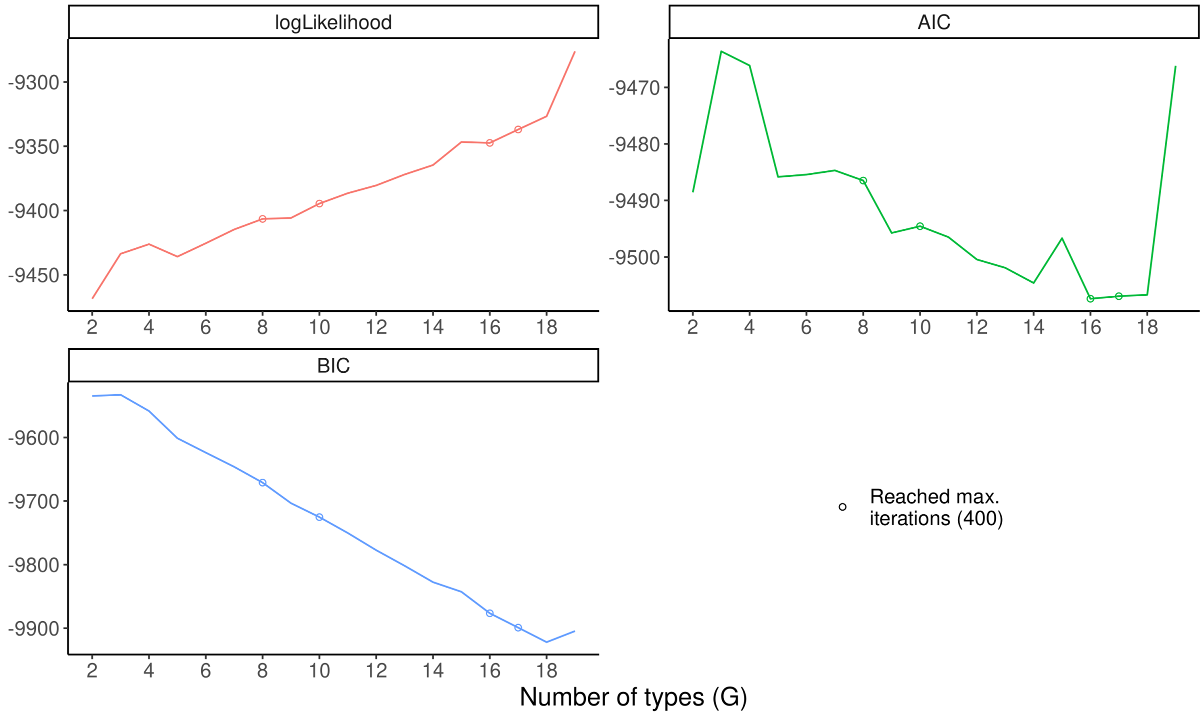


Figure F3: Likelihood criteria: cognitive and non-cognitive measures

(a) Male



(b) Female

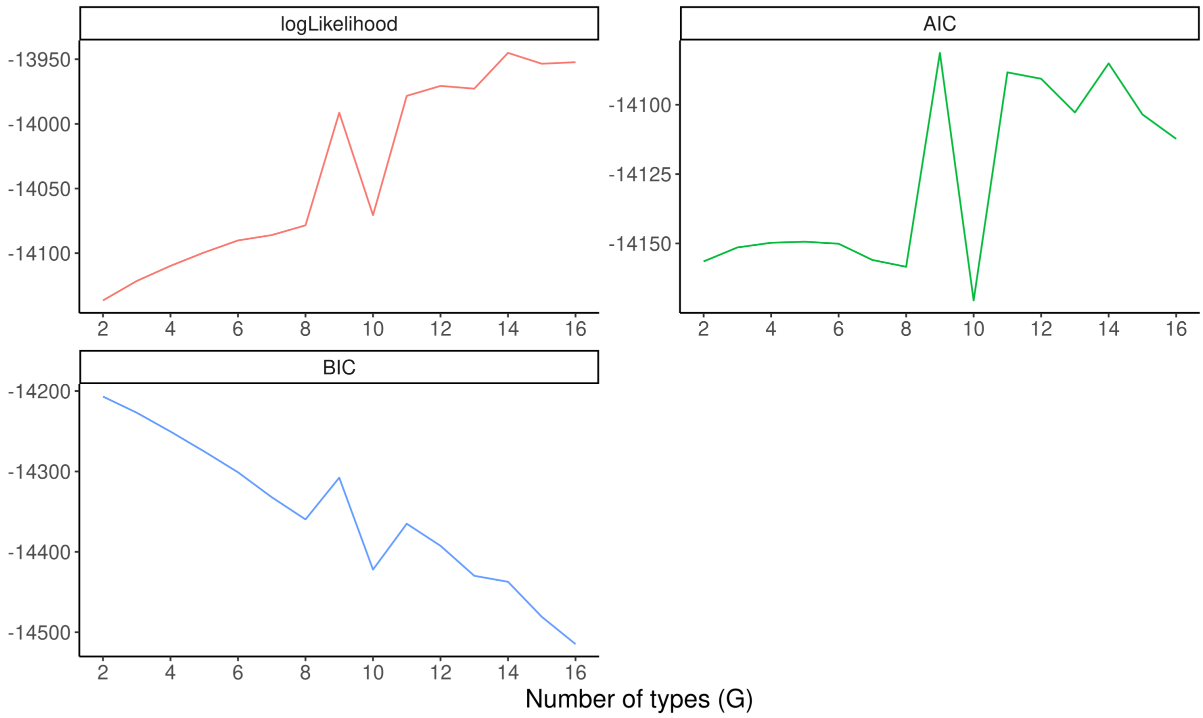
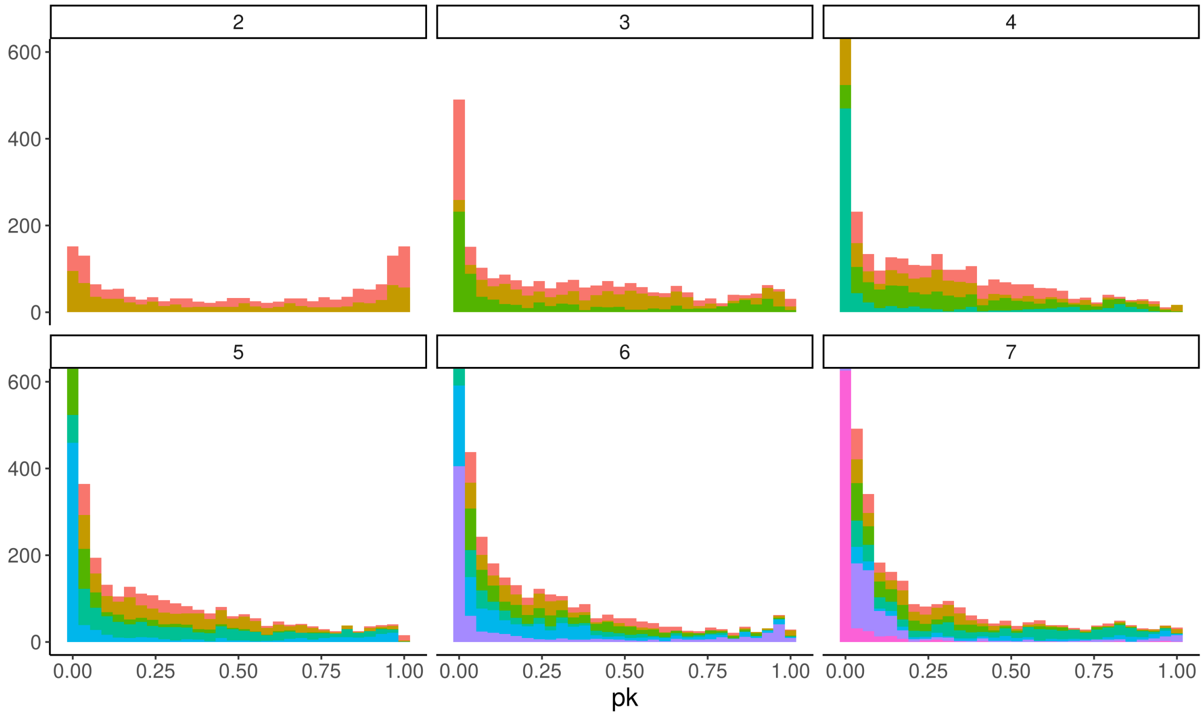


Figure F4: Posterior probabilities: cognitive and non-cognitive measures

(a) Male



(b) Female

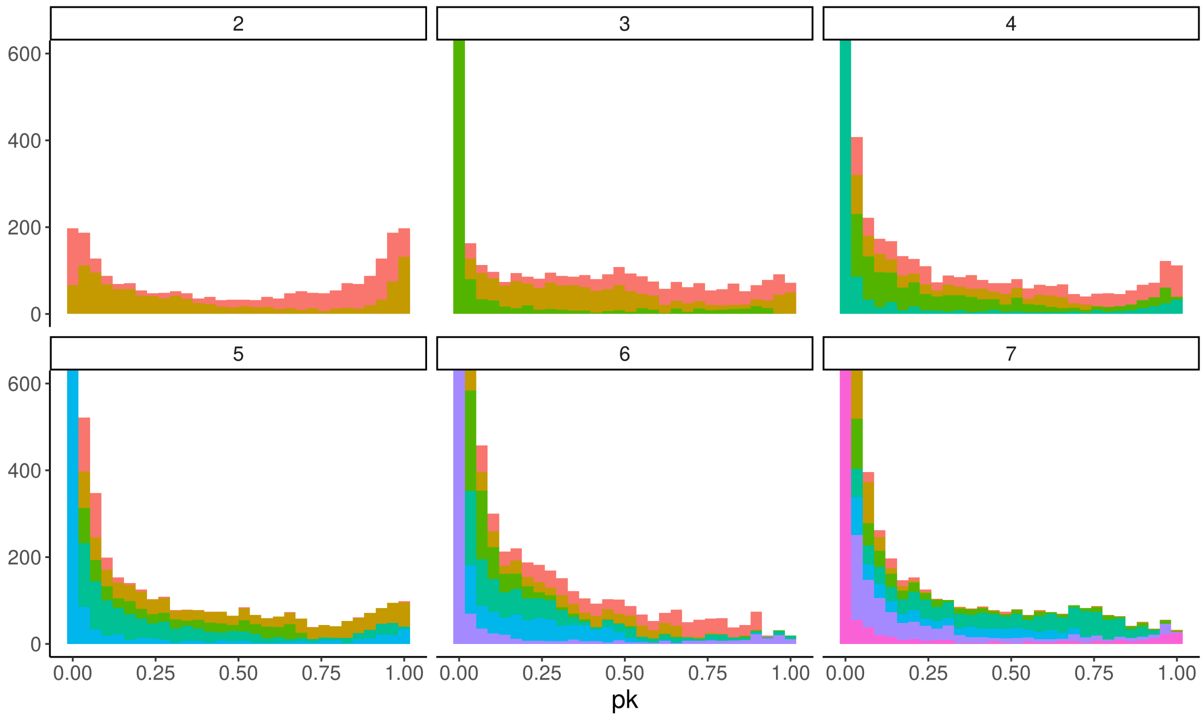


Figure F5: Results across K (single cognitive measure)

(a) Male

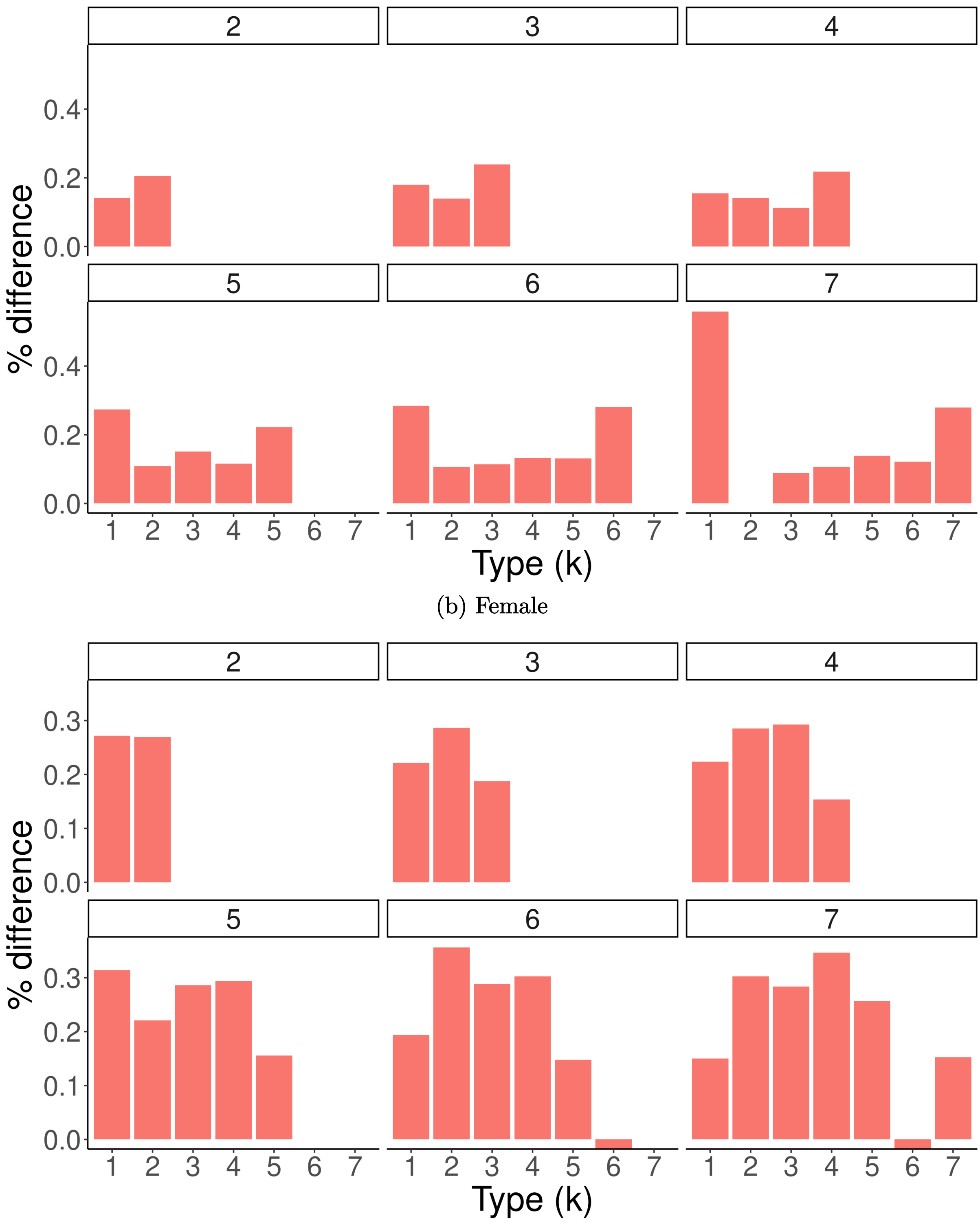
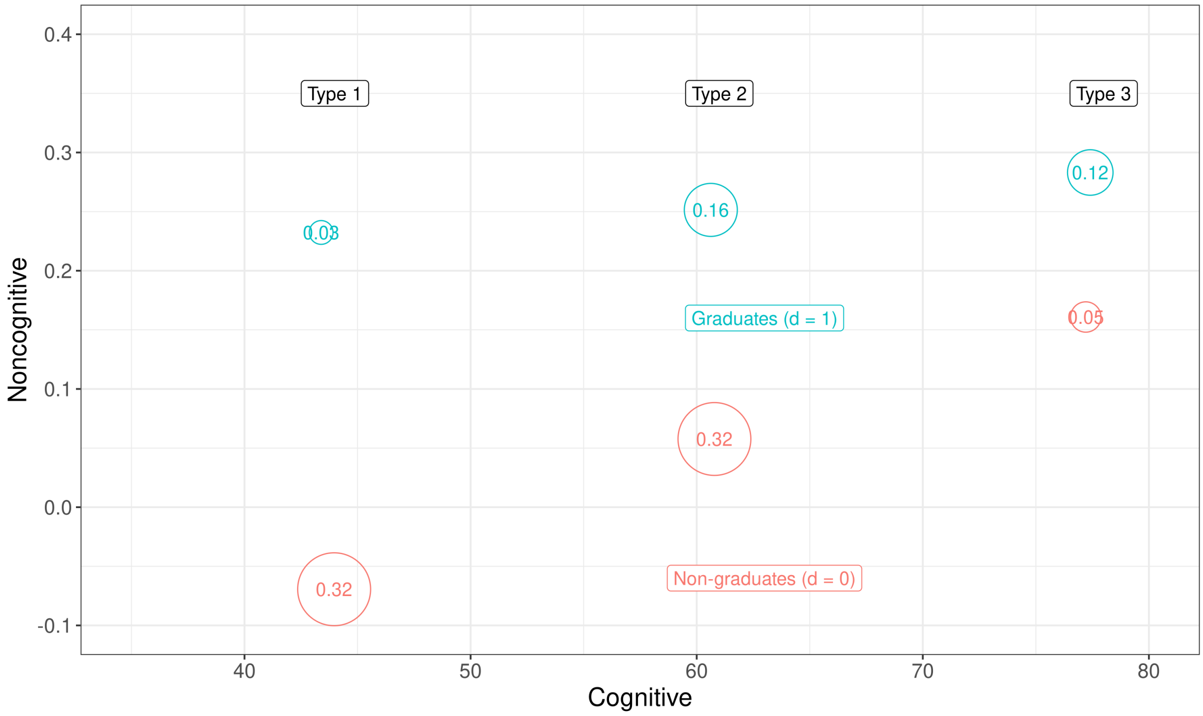
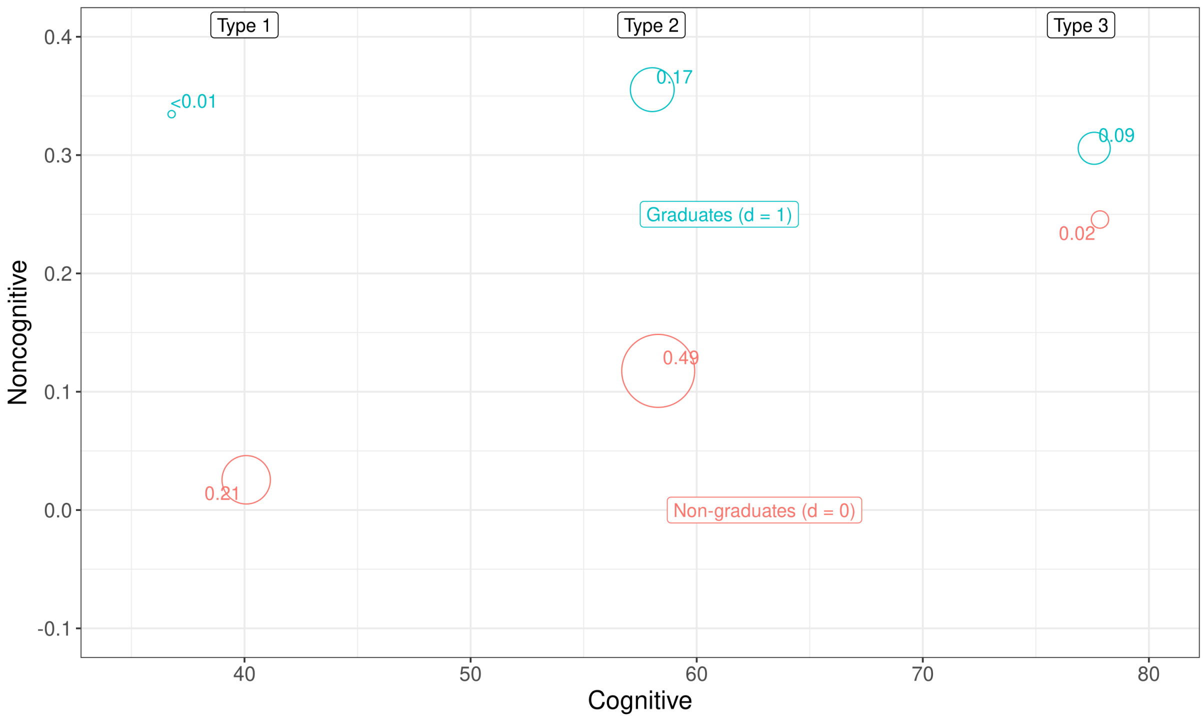


Figure F6: Group sizes and locations in cognitive-noncognitive space ($K = 3$, cognitive measures only)

(a) Male



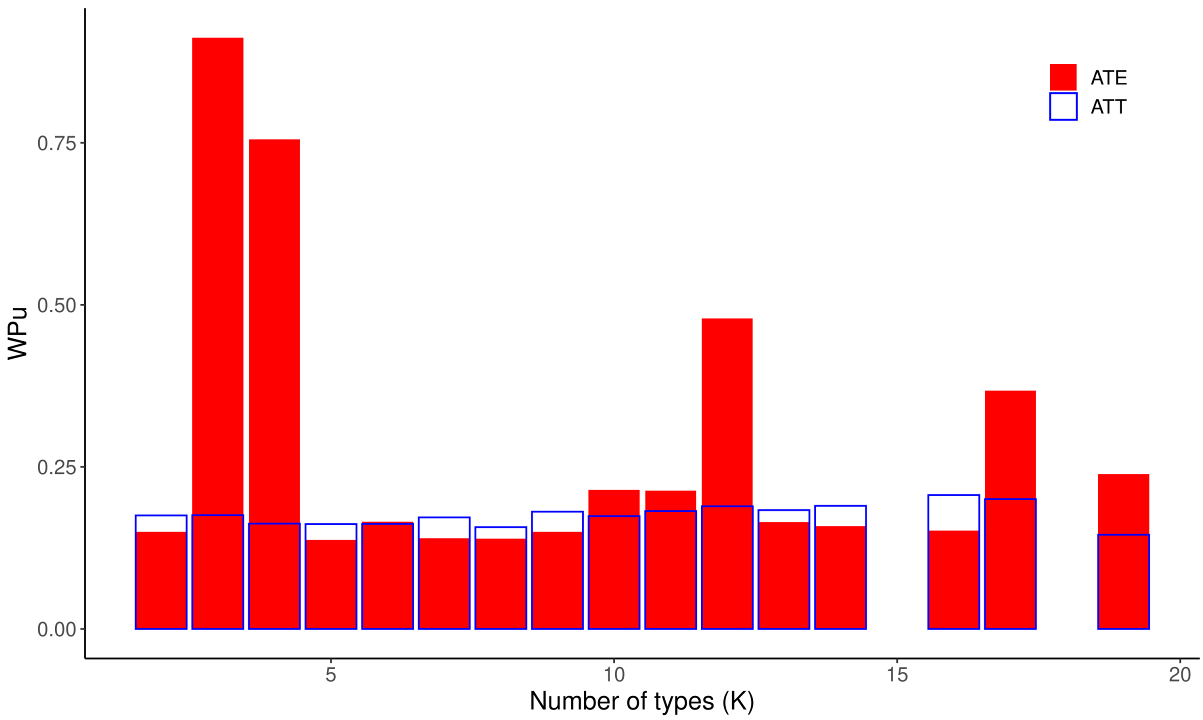
(b) Female



Notes: The above plots display the mean abilities (circle locations) and sizes (circle sizes and labels) for each type, split by gender (panels) and education (colour). Panel (a) contains men, and panel (b) women. Blue circles represent graduates and red non-graduates. The size of each type-education group is labelled, along with each type.

Figure F7: ATEs / ATTs across K (cognitive and noncognitive measures)

(a) Male



(b) Female

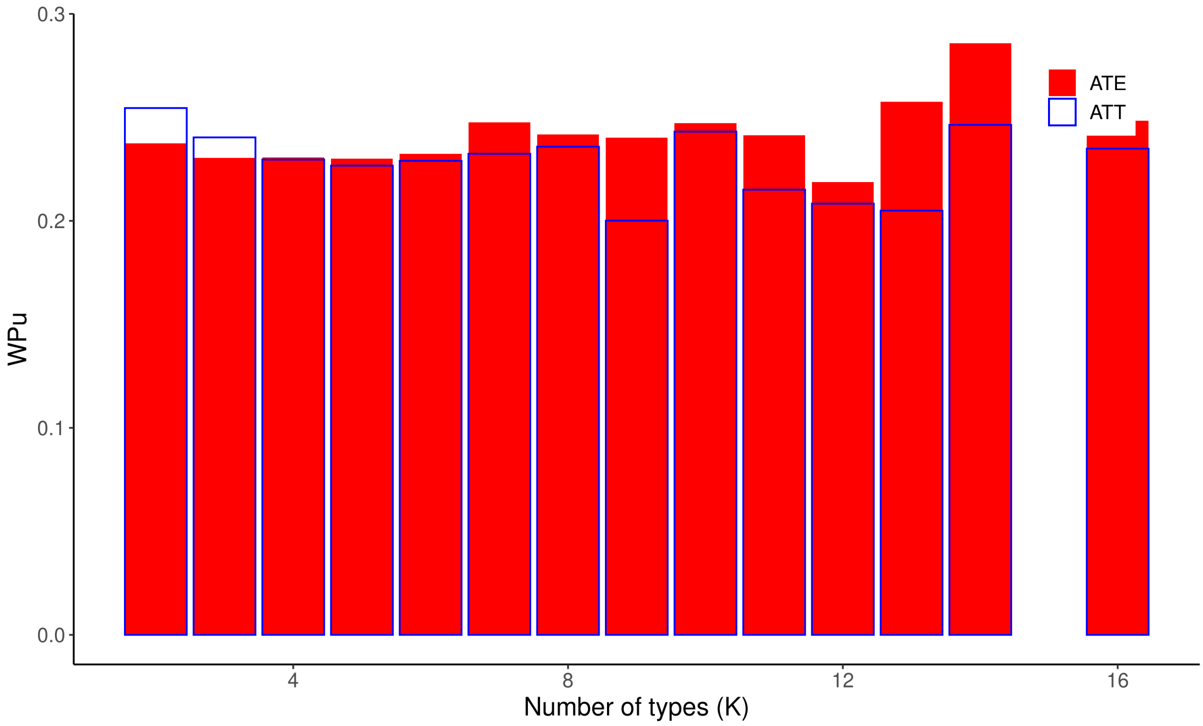


Table F3: $\pi(k, z, d)$ parameter estimates (male, cognitive and non-cognitive)

	$K = 5$									
Type(k) =	1		2		3		4		5	
$d =$	0	1	0	1	0	1	0	1	0	1
Matters very much	0.004	< 0.001	0.045	0.004	0.007	< 0.001	0.031	0.028	< 0.001	0.022
Matters somewhat	0.072	< 0.001	0.119	0.016	0.034	0.004	0.074	0.075	0.015	0.062
Doesn't matter	0.077	< 0.001	0.086	0.016	0.015	0.002	0.072	0.046	0.032	0.042

Table F4: $\pi(k, z, d)$ parameter estimates (female, cognitive and non-cognitive)

	$K = 5$									
Type(k) =	1		2		3		4		5	
$d =$	0	1	0	1	0	1	0	1	0	1
Matters very much	< 0.001	< 0.001	0.080	0.003	0.019	0.008	0.042	0.027	0.009	0.034
Matters somewhat	0.023	< 0.001	0.175	0.007	0.053	0.015	0.065	0.067	0.027	0.046
Doesn't matter	< 0.001	< 0.001	0.161	0.006	0.041	0.008	0.028	0.026	0.007	0.021

where w_i is log weekly wage, d_i is an indicator for univeristy attendance, M_i^C and M_i^N are cognitive and non-cognitive test scores, X_i contains controls for parental income, location type (city/town/countryside), region, and whether the young person is white, and ε_i is a random error term.

We split the sample by gender and present the results for men in panel (a) and women in panel (b). The first column of table F5 presents the results from the most basic specification, an OLS regression log wages on the degree indicator without any controls. Moving across the columns we add controls to the specification, starting with cognitive in the second column, and non-cognitive (column 3), and then all controls (column 4). Adding controls generally decreases the estimates of the returns to a degree, as one might expect given that wage and university attendance are both positively correlated with prior ability. There is one exception: the coefficient on university attendance when all controls are included for males is larger than with just cognitive and non-cognitive test scores. Finally we use 2SLS with the desire to leave home as an instrument for university attendance, first without (column 5) and then with controls (column 6). The 2SLS estimates are slightly larger than our preferred OLS estimates for men, and much larger for women, suggesting either the strong exclusion restriction required for 2SLS does not hold, or the *compliers* who are induced to attend university by the instrument have unusually high returns (interpreting our 2SLS estimate as a LATE). Recall our main analysis does not require the same exogeneity of the instrument as 2SLS.

Our estimates are broadly in line with previous estimates of the returns to university from the UK during this period. Blundell et al. (2000) estimate a similar equation using OLS with detailed controls on data from a UK cohort born 12 years earlier (in 1958), and using wages observed later in the life-cycle at age 33. They estimate returns of around

17% for men and 37% for women. We will return to our OLS and 2SLS estimates in section 2.5 when we use our framework to decompose these estimates using the formulas in section 2.C.

Table F5: OLS and 2SLS estimates of the wage returns to a degree

(a) Male

	Dependent variable: log weekly wage					
	(1)	(2)	(3)	(4)	(5)	(6)
Degree	0.220***	0.181***	0.170***	0.178***	0.207	0.248
	(0.038)	(0.041)	(0.041)	(0.054)	(0.470)	(0.582)
Cognitive		0.003***	0.003**	0.002		0.002
		(0.001)	(0.001)	(0.002)		(0.006)
Non-cognitive			0.057*	0.075*		0.046
			(0.033)	(0.045)		(0.083)
Add. controls				✓		
Instrument					✓	✓
Observations	745	745	745	514	745	745
R^2	0.042	0.052	0.056	0.096	0.042	0.052
Adjusted R^2	0.041	0.050	0.052	0.041	0.041	0.048
Residual se	0.487	0.485	0.484	0.510	0.487	0.486

(b) Female

	Dependent variable: log weekly wage					
	(1)	(2)	(3)	(4)	(5)	(6)
Degree	0.325***	0.291***	0.277***	0.247***	0.530	0.471
	(0.033)	(0.035)	(0.035)	(0.043)	(0.338)	(0.465)
Cognitive		0.003***	0.003***	0.004***		0.001
		(0.001)	(0.001)	(0.001)		(0.004)
Non-cognitive			0.068***	0.034		0.047
			(0.025)	(0.030)		(0.056)
Add. controls				✓		
Instrument					✓	✓
Observations	1,131	1,131	1,131	809	1,131	1,131
R^2	0.078	0.086	0.092	0.150	0.047	0.067
Adjusted R^2	0.078	0.084	0.089	0.118	0.046	0.065
Residual se	0.494	0.492	0.491	0.481	0.502	0.497

Notes: *p<0.1; **p<0.05; ***p<0.01. Specification (1) regresses log-wage on an indicator for a degree and a constant. (2) and (3) include cognitive and noncognitive measures. Then (4) also includes parental income, location type (city/town/countryside), region, and whether the young person is white. Columns (5) and (6) instrument the degree indicator with our instrument.

Chapter 3

A Nonparametric Finite Mixture Approach to Difference-in-Difference Estimation, with an Application to On-the-job Training and Wages

with Robert GARY-BOBO, Julie PERNAUDET, and Jean-Marc ROBIN

Abstract

We develop a finite-mixture framework for nonparametric difference-in-difference analysis with unobserved heterogeneity correlating treatment and outcome. Our framework includes an instrumental variable for the treatment, and we demonstrate that this allows us to relax the common-trend assumption. Outcomes can be modeled as first-order Markovian, provided at least 2 post-treatment observations of the outcome are available. We provide a nonparametric identification proof. We apply our framework to evaluate the effect of on-the-job training on wages, using novel French linked employee-employer data. Estimating our model using an EM-algorithm, we find small ATEs and ATTs on hourly wages, around 1%.

3.1 Introduction

Job training, adult learning, continuous education, or continuing vocational training are just different names for the same thing: the process of (re)training adult workers throughout their working life. A recent OECD report (OECD, 2021) points at large differences in access to training across workers, with the workers most affected by automation being under-represented in training (the unemployed, low-skilled and low-wage workers, and those at smaller firms). Thus, it does not seem that adult learning is attractive to those who need it most.

This could be due to low expectations. The OECD's Priorities for Adult Learning dashboard (PAL dashboard) measures the perceived impact of training by workers through a survey looking at self-reported satisfaction, acquired skills, employment outcomes and wages. In certain countries, such as most Northern European countries (including France and Germany), Japan and Korea, the perceived impact of job training is rather low. For example, France scores around 40%, while the score ranges as low as 20% in the Netherlands and as high as 85% in Chile.

In another OECD research paper on training — their ubiquity showing how important the question of adult education and training has become for society in recent years — Fialho et al. (2019) provide the most recent and exhaustive evaluation of the different forms of adult learning — informal (on the job), non-formal and formal (depending on whether the institution providing training is public or not) — for various countries.[1] The effect of non-formal training on wages is estimated between 13% and 30%, with and without controls. When a control-function estimator is used, the estimated effect of training remains high, around 11% on average, but with a wide range across countries.

Such large returns to training are at odds with the perceived impact of training. Could selection, unobserved heterogeneity, and endogenous treatment lead to such important biases that standard statistical evaluation methods (instrumental variables, selection models, etc.) are unable to provide reliable estimates? Or is the perception just wrong? Our paper makes original contributions to this question, both empirically and methodologically. First, we use a unique survey of (non-formal) adult learning matched with administrative data on wages to precisely evaluate the impact of adult training on wages. Unlike most of the empirical literature on adult training, we use panel data, which allows us to estimate difference-in-difference effects, as we observe individual wages in 2013, training occurs in 2014, and we have post-treatment wage observations in 2014 and 2015. Second, inspired by the work of de Chaisemartin and d'Haultfoeuille (2020) who forcefully dispute the assumption of homogeneous treatments of standard difference-in-difference models, we allow for unobserved heterogeneity in treatment effects and, at the

[1] They use data from various sources: PIAAC, already mentioned, the European Adult Education Survey (EU-AES), the European Continuing Vocational Training Survey (EU-CVTS), and the OECD Structural Analysis database (STAN), providing firm output data.

same time, relax the usual common trend assumption by making it necessary to hold only conditional on unobserved heterogeneity.

Our model offers solutions to a number of issues. Most difference-in-difference theoretical frameworks consider repeated cross-sections instead of true panel data, despite the wide availability and usage of panel data in practice.[2] They all make a common trend assumption. That is, conditional on a set of observed characteristics, the change in the counterfactual outcome in the untreated state must be independent of whether the individual is actually in the treatment group or in the control group. In our setup, the common trend assumption need only hold conditional on *unobserved* types.

The availability of repeated observations of the outcome variable is crucial for identification, all the more so depending on the assumed dynamics. The benefits of panel data in difference-in-difference contexts are studied in Bonhomme and Sauder (2011); Freyaldenhoven et al. (2019) and in Callaway and Li (2019); Li and Li (2019); Sant'Anna and Zhao (2020). These papers maintain a common trend assumption, except for the first one. As far as we know, Bonhomme and Sauder (2011) is the only paper that replaces a standard common trend assumption by a structural assumption on the way unobserved heterogeneity determines outcomes. Specifically, they assume a linear factor structure and solve the (semiparametric) identification problem using nonparametric deconvolution techniques.[3] Freyaldenhoven et al. (2019) share the factor structure of Bonhomme and Sauder's framework and some identification ideas.[4] We depart from the linear factor structure, allowing for instance unobserved heterogeneity to condition outcome variances. Some restriction on the distribution of latent types is however necessary. We assume that there exist a finite number of groups and that outcomes and treatments are drawn from a distribution that is specific to each group. By giving up continuity, we gain more flexibility and also a simpler method of identification based on standard matrix algebra.[5]

Lastly, panel data is necessary but not sufficient for identification. We also need an instrument for training, as in the seminal work of Heckman and Robb (1985); Imbens and Angrist (1994); Heckman et al. (1997, 1998). Our instrument is whether the worker has received information about training offers. It is a variable that affects the treatment

[2]See Heckman et al., 1998; Abadie et al., 2002; Chernozhukov and Hansen, 2005; Abadie, 2005; Athey and Imbens, 2006; D'Haultfoeuille et al., 2021. A recent sub-field of this literature considers staggered treatments (de Chaisemartin and d'Haultfoeuille, 2020; Callaway and Sant'Anna, 2020). We abstract from this type of dynamic treatments, a problem that we would also face with a longer observation period.

[3]More precisely, the special case studied in Section II.B does satisfy the common trend assumption, since, by taking differences in outcomes, the fixed effect disappears. In Section II.C, they allow for different factor loadings on the latent factor, but these factor loadings are assumed independent of the treatment, which comes close to a common trend assumption.

[4]The pre-treatment periods of Freyaldenhoven *et al.* play a similar role as the "instrument" of Bonhomme and Sauder, as far as the identification of factor loadings is concerned.

[5]As for example in Hu and Schennach (2008); Allman et al. (2009); Kasahara and Shimotsu (2009); Hu and Shum (2012); Shiu and Hu (2013); Henry et al. (2014); Hu (2015); Sasaki (2015); Bonhomme et al. (2016a,b, 2017a,b, 2019).

probability for each type, potentially in a different way. It makes sense to expect monotonicity to hold in our application — more information on training possibilities increases the probability of training — but it is not needed here. In addition, the potential outcomes are assumed to be conditionally independent of the instrument and the treatment, but only conditional on the unobserved type. The "instrument" (in this weaker sense) allows us to identify the model nonparametrically.

Under these assumptions about discrete heterogeneity and the instrument, we prove identification of Average Treatment Effects (ATEs) that are conditional on the unobserved types, as well as heterogeneous treatment probabilities. We also show that each standard difference-in-difference estimator obtained by applying OLS or IV estimation procedures to the wage panel equations is the sum of different weighted means of conditional average treatment effects (ATT and LATE), plus a bias reflecting the absence of common trend.

After proving identification, we estimate a flexible parametric specification using the (sequential) Expectation-Maximization (EM) algorithm. Standard errors are obtained by Bootstrap. The results show that treatment effects vary with type, but once aggregated they are very small and insignificant. All three ways of aggregating conditional ATEs (aggregate ATE, ATT and LATE) yield similar estimates of around 1%. We conclude that on-the-job training has no or a very limited effect on wages. The biases resulting from heterogeneous trends are found to be of a similar order of magnitude to the aggregate treatment effects. We also find a sizable share of the bias on the IV estimator of around 1%-2% that reflect small sample deviations from assumed restrictions in the population.

The rest of this paper proceeds as follows. We first end the introduction by a discussion of the literature on training. Then, Section 3.2 presents the model, the associated nonparametric identification result as well as the links between our model's estimates and ATT, ATE and other parameters of interest. Section 3.3 describes our dataset and presents a preliminary econometric analysis using standard econometric methods. Section 4 presents and discusses the results of the estimation of our model. We conclude in Section 5.

Literature on training. The literature on the effect of training (and active labor market programs) is huge. The estimated impacts of training on wages and productivity are generally found to be positive; the effects on the risk of unemployment are often ambiguous. Before Fialho et al. (2019), several other authors have reviewed this literature (see Heckman et al. (1999); McCall et al. (2016) and the meta-analyses of Card et al. (2010, 2018) and Haelermans and Borghans (2012)). See also the classic paper by LaLonde (1986).

Many of the contributions devoted to training programs are based on non-experimental data with a panel structure and rely on fixed-effects estimators. Fixed-effects approaches are used in the pioneering work of Ashenfelter (1978), in the contributions of (among

many others) Lynch (1992), on NLSY data; Booth (1993); Blundell et al. (1999), both on British data; Krueger and Rouse (1998), on American firm-level data; Pischke (2001), on German GSOEP data; Schoene (2004), on Norwegian data.

Few papers rely on instrumental variables, maybe because it is difficult to find convincing instruments for participation in training programs (yet, see Bartel (1995); Parent (1999); Abadie et al. (2002)). Some contributions controlled for selection in training using Heckman's two-stage estimator (*e.g.* LaLonde (1986); Booth (1993); Goux and Maurin (2000)). A behavioral approach to training participation is explored in Caliendo et al. (2016). Other contributions use matching estimators (Brodaty et al., 2001; Gerfin and Lechner, 2002; Kluve et al., 2012).

A number of recent papers follow Abadie et al. (2002) and use randomized trials; see *e.g.* Lee (2009), Attanasio et al. (2011); Grip and Sauermann (2012); Ba et al. (2017), Sandvik et al. (2021). The importance of the comparison group construction is illustrated by Leuven and Oosterbeek (2008). They narrow down their comparison group to pick only "workers who [were] willing to undertake training and whose employers [were] prepared to provide it, but did not attend the training course they wanted, due to some random event" (Leuven and Oosterbeek, 2008, p. 426). This strict choice of comparison group reduces the estimated coefficient on training to almost zero, down from between 5–15% for less restrictive choices.[6]

Most papers consider the impact of training on wages *and* productivity. Human capital theory suggests that, under conditions of perfect competition, employers should refuse to pay for training. At least, they would refuse to finance general training, that is typically portable, and would allow workers to quit the firm and find a job with a higher wage. But under imperfectly competitive conditions, in particular, under asymmetric information about workers' abilities, it can be shown that the firm should be willing, either to subsidize training, or to share the benefits of training with the worker, (see Acemoglu and Pischke, 1998, 1999). A number of papers use wage equations and production functions to test this prediction and do indeed find positive effects on both productivity and wages.[7]

There also exists a literature on transition and duration models, studying the effects of training on the duration of employment and unemployment spells (see Ridder (1986), on Dutch data; Gritz (1993), on NLSY data; Bonnal et al. (1997), on French data; Crepon et al. (2009), using methods developed in Abbring and Berg (2003)).

Finally, an important question is to assess the importance and effects of unobserved heterogeneity, as well as the dynamic structure of the treatment effects of training (for recent progress on these two fronts, see Rodriguez et al. (2018)).

[6] On this point, see also Sandvik et al. (2021).

[7] See Ballot et al. (2006), Dearden et al. (2006); Konings and Vanormelingen (2015).

3.2 The model

We study a population of N workers indexed by i. The worker's nominal hourly wage (in logs) is denoted w_{it} and is observed at the end of three consecutive years indexed by $t = 1, 2, 3$. Some workers engage in a training session after the first wage observation, in which case $d_i = 1$, and $d_i = 0$ otherwise. Wage w_{i1} is observed before training, and w_{i2}, w_{i3} are observed after training (if training takes place at all). Our goal is to measure the causal impact of training on the wages in periods $t = 2, 3$. In this empirical application, we assume that treatment d_i is a binary variable, although the model and the proof of identification encompass the case of a treatment variable with any finite number of values. Specifically, we could allow for different types of training, by duration for example.

We single out from all potential control variables a variable $z_i \in \{0, 1\}$, indicating if the worker reports receiving information about the availability of training sessions through any of the following channels: hierarchy, human resources, coworkers, or unions. This variable will be used as an instrument for the selection into treatment.

We assume that workers can be clustered into a finite number H of unobserved groups: $h \in \{1, ..., H\}$. The distribution of all variables w_{it}, z_i and d_i, including the instrument, potentially varies across latent groups. In our application, we do not use any control variables. However, semiparametric versions of our model can easily be constructed and estimated, at the cost of restrictions on the interaction between observed and unobserved characteristics. We classify workers from observations (w_{it}, d_i, z_i) and examine the correlations between the estimated classification h and a set of available controls, ex post.

We start by making the following assumption on wages.

Assumption 7 (Wage process)**.** *The wage process is first-order Markov and independent of the instrument given type and treatment.*

First, we allow for first-order autoregressive dependence in wages after controlling for unobserved heterogeneity. The static case is a particular case where w_{i2} is a constant. Second, the variable z_i is a valid instrument for training insofar as it does not affect wages once heterogeneity and training are controlled for. This exclusion restriction is really the most fundamental one in our framework.

Let $f_t(w_{it}|h, d)$ denote the density function for the marginal distribution of wages w_{it} given treatment and type. Let $f_{t|s}(w_{it}|w_{is}, h, d)$ denote the density function for the conditional distribution of w_{it} given w_{is} (we use $s = t \pm 1$). Let $\pi(h, z, d)$ be the probability mass of workers of type $h \in \{1, ..., H\}$, with values of the instrument $z \in \{0, 1\}$ and of treatment $d \in \{0, 1\}$. Let $\mathcal{W}_2(h, d)$ be the support of $f_2(w_2|h, d)$, and let $\mathcal{W}_2(d) = \bigcap_{h=1}^{H} \mathcal{W}_2(h, d)$ be the common support.

Our framework relates to the standard difference-in-differences model, since we compare post- and pre-treatment wages. It can also be interpreted as a version of the Roy model used by Heckman and Vytlacil in numerous papers including Heckman and Vytlacil

(2005); Carneiro and Lee (2009); Carneiro et al. (2010, 2011). The main difference is that we explicitly model the dependence between error terms via the latent factor h, which is moreover assumed discrete. Specifically, a possible interpretation of our model combines an outcome and a choice (or selection) equation as follows. Let $y(0)$, $y(1)$ denote the potential outcomes (i.e., post-treatment wages) for $d=0$ or $d=1$ and let $c(h,z)+v$ be a random training cost depending on h and z. Then, a standard Roy model would have $d=1$ if and only if the expected return is greater than the cost:

$$\mathbb{E}[y(1)-y(0)\,|\,h] \geq c(h,z)+v.$$

3.2.1 Identification

In this section, we describe the conditions under which our model is identified.

All the relevant information is in the likelihood of the information available at the individual level, namely the instrument z, the treatment d, and the three wages w_1 (before treatment) and w_2, w_3 (after treatment):

$$p(z,d,w_1,w_2,w_3) = \sum_h \pi(h,z,d)\, f_2(w_2|h,d)\, f_{1|2}(w_1|w_2,h,d)\, f_{3|2}(w_3|w_2,h,d), \tag{3.1}$$

where, by Bayes' rule,

$$f_{1|2}(w_1|w_2,h,d) = \frac{f_1(w_1|h)\, f_{2|1}(w_2|w_1,h,d)}{f_2(w_2|h,d)}.$$

We show that all the components of the right-hand side of equation (3.1) are identified under the following assumptions.

Assumption 8 (Overlap). *For all d, $\pi(h,0,d) \neq 0$ for all h, and $\mathcal{W}_2(d) \neq \emptyset$.*

Assumption 8 is standard and means that workers of all types have a positive probability of being both treated and non treated for at least one instrument value, arbitrarily set equal to zero. Moreover, some second period wages can be realized with positive probability irrespective of training. This assumption was implicit to the above identifying restriction (3.1). Otherwise, the sum is only over those types h such that $f_2(w_2|h,d) \neq 0$. In the static case, this assumption is automatically satisfied.

Assumption 9 (Linear independence). *For all d and all w_2 in $\mathcal{W}_2(d)$, $[f_{t|2}(w_t|w_2,1,d),\cdots, f_{t|2}(w_t|w_2,H,d)]$, $t=1,3$, are linearly independent systems.*

Any latent type such that its conditional wage distribution can be replicated as a linear combination of the other types' distributions cannot be separately identified from the other types. This is the nonparametric equivalent of the standard rank condition for OLS.

Assumption 10 (First stage). *For all d, $\dfrac{\pi(h,1,d)}{\pi(h,0,d)} \neq \dfrac{\pi(h',1,d)}{\pi(h',0,d)}$ for all $h \neq h'$.*

Assumption 10 requires different exposures to the instrument for all types whatever the treatment. There are two ways to think about this assumption. First, z could be perfectly randomized and independent of heterogeneity h (say, $\pi(z,h) = \pi(z)\pi(h)$). In this case, Assumption 10 requires strong dependence of conditional treatment probability $\pi(d|z,h)$ on both z *and* h. Alternatively, the instrument could be weak as long as it is dependent on unobserved type, i.e. $\pi(d|z,h) = \pi(d|h)$ and $\pi(z,h) \neq \pi(z)\pi(h)$.

Assumptions 7, 8, 9 and 10 are the main conditions for recovering the underlying structure, separately, given d and $w_2 \in \mathcal{W}_2(d)$. More precisely, under these assumptions, we show that the three functional components: $\pi(h,z,d)\, f_2(w_2|h,d)$, $f_{1|2}(w_1|w_2,h,d)$ and $f_{3|2}(w_3|w_2,h,d)$ are separately identified for all h,d and $w_2 \in \mathcal{W}_2(d)$. The odds ratios $\frac{\pi(h,1,d)}{\pi(h,0,d)}$ being independent of wages, they also allow to identify a common labeling of the latent groups over wages $w_2 \in \mathcal{W}_2(d)$.

Assumption 11 (Identical supports). *For all d, distributions $f_2(w_2|h,d)$, $h = 1,...,H$, have identical support: $\mathcal{W}_2(h,d) = \mathcal{W}_2(d)$ for all h.*

In practice, if the common support $\mathcal{W}_2(d)$ is not the full support for all types, then identification, at least with the method of this paper, fails.[8] Under Assumption (11) we can further identify $\pi(h,z,d)$ from $f_2(w_2|h,d)$, and also $f_1(w_1|h,d)$.

Assumption 12 (Predetermination). *For all types h, $f_1(w_1|h,d) = f_1(w_1|h)$.*

This assumption implies that the first period wages are predetermined. They are independent of treatment given latent type. This is a natural assumption that allows to label groups identically across treatments. If first period wages were not predetermined, then, other assumptions could ensure a common group labeling. For example, we could assume that the rank of mean wages by type is invariant with respect to treatment.

We finally add the following assumption.

Assumption 13 (Discrete wages). *Wage distributions are discrete.*

Assumption 13 is added without much loss of generality. It will spare us a number of technicalities. For continuously distributed wages, we could project the function $(w_1,w_2,w_3) \mapsto P(z,d,w_1,w_2,w_3)$ on a functional basis, and it would be relatively straightforward to adapt the identification argument.

We can now state our main identification theorem.

Theorem (Identification). *Under Assumptions (7)-(13), the model parameters $\pi(h,z,d)$, $f_1(w_1|h)$, $f_{2|1}(w_2|w_1,h,d)$, and $f_{3|2}(w_3|w_2,h,d)$ are identified.*

[8] If only because there is no way to say a priori which wages have non zero probability for some types and not for the other types.

The proof of the identification theorem is in Appendix A. It uses similar matrix factoring arguments (singular value decomposition) as in Kasahara and Shimotsu (2009); Hu and Shum (2012); Bonhomme et al. (2016a,b, 2017a).[9]

We are now equipped with a nonparametric identification result and we can safely develop a method to estimate our model. Our estimation method is described in Sections 3.3.3 and 3.3.4 below. Although the identification proof is constructive, it leads to complicated estimating equations that do not use all the available information. This is why we prefer, for estimation, to use maximum likelihood and a parametric version of the model.[10]

3.2.2 Treatment effects and usual estimators

Before turning to the estimation procedure and our application to training and wages, we discuss the definition of policy-relevant parameters in our framework. We mainly compare the treatment effects with the usual estimators of applied econometrics, such as OLS and IV estimators.

Let $y(0)$ and $y(1)$ denote the *counterfactual outcom*es. In our application, it can be the wages in period $t=2$ or $t=3$ of untrained and trained workers, or the wage changes between $t=2,3$ and $t=1$ given training. Hence, our discussion will encompass both static and dynamic experiments (yet not staggered treatments). Note also that, in our setup, counterfactual outcomes $y(0)$ and $y(1)$ satisfy the conditional independence assumption:

$$y(0), y(1) \perp\!\!\!\perp d, z \,|\, h. \tag{3.2}$$

The difficulty here is that the conditioning variable h is not observed.

Define the observed outcome $y = dy(1) + (1-d)y(0)$. We now define and derive the *Average Treatment Effect* (ATE) and the *Average Treatment Effect on the Treated* (ATT). Then we consider OLS and IV estimators.

ATE. We define a *conditional* Average Treatment Effect given type h as follows,

$$ATE(h) = \mathbb{E}[y(1) - y(0)|h] = \mu(h,1) - \mu(h,0),$$

where $\mu(h,d) = \mathbb{E}[y(d)|h]$. The unconditional ATE is simply the average over types $h = 1, \ldots, H$ of the conditional ATEs, that is,

$$ATE = \sum_h \pi(h)\, ATE(h),$$

[9] Hu and Shum (2012) consider very general mixtures of Markovian processes. By avoiding sophisticated operator algebra, thanks to discrete wages, we can propose a more straightforward proof.

[10] Our parametric version could be made arbitrarily flexible, but the data that we use would not support the estimation of a complicated specification with a large number of parameters.

where $\pi(h) = \sum_{z,d} \pi(h,z,d)$ is the population share of type-h workers.

ATT. Under the above conditional independence assumption,

$$ATT(h) = \mathbb{E}[y(1) - y(0)|h, d = 1] = ATE(h).$$

The ATT is thus the average value of the conditional treatment effect $ATE(h)$ over the treated individuals:

$$ATT = \mathbb{E}[y(1) - y(0)|d = 1] = \sum_h \pi(h|d = 1) ATE(h),$$

with $\pi(h|d) = \sum_z \pi(h,z|d)$ and for $d = 0,1$,

$$\pi(h,z|d) = \frac{\pi(h,z,d)}{\sum_{h,z} \pi(h,z,d)}.$$

OLS and DiD. Now, we study the OLS estimator of the impact of treatment on the outcome. The *difference-in-difference* (DiD) estimator is the OLS estimator when the outcomes $y(1)$, $y(0)$ are defined as wage changes between before and after the treatment's application.

We have

$$\begin{aligned} b_{OLS} &= \frac{\mathrm{Cov}(y,d)}{\mathrm{Var}(d)} = \mathbb{E}[y(1)|d = 1] - \mathbb{E}[y(0)|d = 0] \\ &= \sum_h \pi(h|d = 1)\,\mu(h,1) - \sum_h \pi(h|d = 0)\,\mu(h,0) \\ &= ATT + B_{OLS}, \end{aligned}$$

where $\pi(h|d) = \sum_z \pi(h,z|d)$ and B_{OLS} is the bias, defined as

$$B_{OLS} = \sum_h [\pi(h|d = 1) - \pi(h|d = 0)]\,\mu(h,0).$$

Hence, the OLS estimator is an unbiased estimator of ATT if

1. $\pi(h|d = 1) = \pi(h|d = 0)$ for all types h; or

2. $\mu(h,0) = \mu(1,0)$ for all h.

These restrictions will not hold in general as we expect neither the decision to treat, nor the outcome levels to be independent of individual types. However, with outcomes defined as wage changes between periods before and after training, assumption 2 is the usual common trend assumption in DiD setups: the expected change in the outcome, before and after treatment, is independent of the group. Hence, for levels, we shall refer to B_{OLS}

simply as the "heterogeneity" bias. For differences, we will call B_{OLS} the "heterogeneous trend" bias.

Lastly, the sign of the bias is unknown *a priori.* However, imagine that good types, with higher pre-treatment wages (and wage growth), also have a higher probability of benefiting from training. Then, we expect the OLS estimator (or the DiD) to be biased upward vis-a-vis the ATT. One can find a similar discussion in Carneiro et al. (2011).

IV and LATE. Finally, the IV estimator of the regression of y on d, using z as instrument can be expressed as follows,

$$b_{IV} = \frac{\mathrm{Cov}(y,z)}{\mathrm{Cov}(d,z)} = \frac{\mathbb{E}(y|z=1) - \mathbb{E}(y|z=0)}{\mathbb{E}(d|z=1) - \mathbb{E}(d|z=0)}.$$

First, the denominator of b_{IV} is trivially

$$\mathbb{E}(d|z=1) - \mathbb{E}(d|z=0) = \sum_h \left[\pi(h,d=1|z=1) - \pi(h,d=1|z=0)\right].$$

Second, the numerator can be factored as

$$\begin{aligned}
\mathbb{E}(y|z=1) - \mathbb{E}(y|z=0) &= \sum_h \left[\pi(h,d=1|z=1)\,\mu(h,1) + \pi(h,d=0|z=1)\,\mu(h,0)\right] \\
&\quad - \sum_h \left[\pi(h,d=1|z=0)\,\mu(h,1) + \pi(h,d=0|z=0)\,\mu(h,0)\right] \\
&= \sum_h \left[\pi(h,d=1|z=1) - \pi(h,d=1|z=0)\right]\left[\mu(h,1) - \mu(h,0)\right] \\
&\quad + \sum_h \left[\pi(h|z=1) - \pi(h|z=0)\right]\mu(h,0),
\end{aligned}$$

making use of

$$\pi(h,d|z) = \frac{\pi(h,z,d)}{\sum_{h,d}\pi(h,z,d)} \quad \text{and} \quad \pi(h|z) = \sum_d \pi(h,d|z).$$

Hence,

$$b_{IV} = LATE + B_{IV},$$

where we define

$$LATE = \frac{\sum_h \left[\pi(h,d=1|z=1) - \pi(h,d=1|z=0)\right] ATE(h)}{\sum_h \left[\pi(h,d=1|z=1) - \pi(h,d=1|z=0)\right]} \tag{3.3}$$

and

$$B_{IV} = \frac{\sum_h \left[\pi(h|z=1) - \pi(h|z=0)\right]\mu(h,0)}{\sum_h \left[\pi(h,d=1|z=1) - \pi(h,d=1|z=0)\right]}. \tag{3.4}$$

LATE is a weighted average of conditional ATEs given type. This average is infor-

mative if the weights are uniformly positive or negative, that is, if monotonicity holds (Imbens and Angrist, 1994):

$$\pi(h,d=1|z=1) \geq \pi(h,d=1|z=0).$$

In our setup, it makes sense to think that the probability of training increases if the employer informs its workers about training possibilities. However, our estimator is more generally applicable as we do not need to assume monotonicity in the treatment probability. As in de Chaisemartin and d'Haultfoeuille (2020)'s application to difference-in-difference, we can check whether all weights are of the same sign or not.

The IV estimator is an unbiased estimator of LATE (i.e., $B_{IV}=0$) if

1. $\pi(h|z=1)=\pi(h|z=0)$ for all types h; or
2. $\mu(h,0)=\mu(1,0)$ for all h.

The second restriction has already been discussed in the case of OLS. The first restriction is also similar, although it now relates heterogeneity to the instrument (z) instead of the treatment (d) itself. In our application, the instrument is determined at the firm level. So, it may be correlated with worker types either because of matching — good firm types matching with good worker types — or if employers themselves inform workers about training possibilities in a selective way. In many usual LATE setups, the instrument is not local (a policy designed at some regional level, for example). In which case, the first restriction is also more likely to hold (that is, if individuals do not move in response to the policy). In randomized setups, z is the intention to treat, the random assignment to treatment and is by construction exogenous. Then, treated individuals may comply ($d=1$) or not ($d=0$) with the assignment to treat (*e.g.*, Abadie et al., 2002).

Conclusion. Our setup therefore offers two main advantages: 1) It allows one to identify average treatment effects (and more generally their distribution across latent types) in situations where counterfactual outcomes are heterogeneous. In a difference-in-difference setup, this means that identification does not rest on the common trend assumption. 2) Identification is complete, meaning that all parameters of the structural model are nonparametrically identified. This allows one to identify not only the conditional treatment effects given types, but also the joint distribution of treatment and types. Hence, the weights of marginal treatments effects in OLS and IV estimators can be separately identified, with no need for such assumptions as constant-sign or monotonicity.

3.3 Application: the wage returns to training

3.3.1 The data

We use survey data collected between 2013 and 2015 by Céreq,[11] as part of the DEFIS survey.[12] The survey sampled 4,529 firms with three employees or more from all sectors but agriculture in 2013, and 16,126 workers were subsequently drawn from these firms' employees.[13] The main objective of the survey was to document the use of formal or non-formal adult education by employees, and the effect of this form of learning on work outcomes. Several waves of interviews were — and still have to be — conducted. We use the first wave in this paper, in which employees were interviewed between June and October 2015 about any training sessions that they participated in between January 2014 and the time of the interview. This was done through retrospective questions (such as "Did you hold a full-time or a part-time contract in firm X in the fall of 2013?", or "Since January 2014, did you take part in a training program?").

The responses to the employer survey (in December 2014) and the worker survey (in 2015) are matched with wage data obtained from tax registers, reported by employers to the tax authorities (*Déclarations annuelles de salaires*, DADS) for the ongoing employment spells in December 2013, December 2014 and December 2015.[14] Our definition of the wage is total earnings paid to the worker by the employer in December 2013, 2014 and 2015, net of payroll taxes (but not net of income tax) and divided by the total number of hours worked in that employment in the whole years of 2013, 2014 and 2015. Nearly 80% (12,597/16,126) of workers reported that they were employed by the same firm as in 2013 at the time of the interview in 2015. Greater fractions (89.2% = 12,100/13,562 in 2014 and 85.3% = 11,103/13,014 in 2015) of the wages recorded for 2014 and 2015 were paid by the same employer who paid the wage recorded in 2013. Therefore, a large majority of workers in our data did not move during our period of analysis so we will abstract from worker mobility in this paper.

To give a first overview of the factors affecting the selection into training, we start with a simple comparison of employees who reported at least one training session in 2014 or 2015 with employees who did not declare any training. Among the 16,126 employees surveyed in 2015, 6,349 individuals (39.3%) declared at least one training session, with a majority of them declaring only one session.[15] Table 3.3.1 presents the average characteristics

[11] *Centre d'études et de recherches sur les qualifications* (a French public institution).

[12] *Dispositif d'enquêtes sur les formations et itinéraires des salariés.*

[13] The employees were sampled among the sampled firms' employees, provided that they were employed by their firm in December 2013. The latter sampling is stratified to provide a representative sample of workers

[14] More precisely, the last employment spells of the years 2013, 2014 and 2015, which ends at the end of December for 83% of the workers in 2013, 78% in 2014 and 76% in 2015.

[15] Among the 6,349 employees who received training, 61% declared one session, 26% declared two, 9% declared 3, and less than 4% declared more than 3.

Table 3.3.1: Comparison of trained and untrained workers by baseline characteristics

	All		Stayers	
	Trained	Untrained	Trained	Untrained
Demographics:				
Age (modal group)	40-44	45-49	40-44	45-49
Male	70.7	67.3	74.1	72.9
French	97.0	94.1	98.1	95.8
In couple	74.8	68.4	78.6	74.0
Has children	57.4	49.0	63.1	55.9
Disability	7.2	12.5	6.7	10.1
Previous health problem	3.4	5.7	2.6	4.1
Education:				
Less than high school diploma	28.3	46.1	29.4	46.5
High school diploma	18.5	18.6	18.3	18.1
Trade or vocational degree	20.7	14.9	21.9	16.7
Bachelor's degree	7.9	5.5	6.9	4.7
Master's degree or more	23.9	13.8	23.0	13.1
Occupation:				
Unskilled blue collar	5.9	9.6	5.2	9.0
Skilled worker, technician	18.5	26.2	18.8	26.8
Office worker, public sector employee	21.2	27.9	17.5	24.3
Foreman/Supervisor	13.7	9.9	15.0	11.2
Technician, draftsman, salesman	9.3	6.5	10.0	7.4
Engineer, manager	29.5	15.7	32.7	18.9
Job characteristics:				
Log(hourly wage), 2013 (w_1)	2.7	2.5	2.8	2.6
Log(hourly wage), 2014 (w_2)	2.8	2.6	2.8	2.6
Log(hourly wage), 2015 (w_3)	2.8	2.6	2.8	2.6
Permanent contract	90.0	83.3	98.5	98.3
Full time contract	88.7	80.1	95.9	93.9
Information on training (z)	78.8	62.8	81.7	68.5
Firm characteristics:				
3 to 49 employees	24.0	39.1	21.1	38.2
50 to 249 employees	20.5	21.8	21.4	23.5
250 to 499 employees	9.1	7.2	9.7	7.9
500 to 999 employees	8.6	6.5	8.5	6.8
1000 to 1999 employees	7.4	6.2	7.2	5.4
More than 2000 employees	30.4	19.1	32.1	18.3
Has HR department	89.6	81.5	91.5	81.9
Has individual incentive strategy	72.4	60.0	74.4	61.8
Has collective incentive strategy	78.4	64.5	82.2	68.6
Outsources part of activity	40.6	34.8	41.9	36.5
Number of observations	6343	9783	3467	4066

Notes: "All" refers to the whole sample and "Stayers" refers to the sub-sample of workers who remain employed in the same firm all three years. For all binary variables, the mean is given as a percentage. The bottom row gives the number of workers for all variables except log(hourly wage), where 59 observations are missing wages in 2013, and approx. 3,000 in 2014 and 2015.

of trained and untrained workers in terms of demographics, education, occupation, job and firm characteristics, before any training (situation in the fall of 2013). Statistics are presented both for the overall sample (the two left-hand columns) and the analysis sample (the two right-hand columns). The analysis sample excludes some individuals with extreme wage observations and more importantly, includes only "stayers" — workers who are observed in the same firm in 2013 and in 2015.

All variables in rows are binary, except the age and hourly wage (in logs). Table 3.3.1 suggests that on average, workers who trained between January 2014 and the time of the first interview (between June and October 2015) are more likely to be French, male, living as a couple, and to have children (even controlling for age) compared to workers who did not train. They also tend to be more educated, most of them having post-secondary degrees. They occupy more skilled jobs, they have higher salaries, and they are more likely to hold full-time and permanent contracts. They are also more likely to receive information on training (our instrument). Using the employer survey, we also find that trained workers are on average in bigger firms, that are more likely to have human resource staff. Overall, more advantaged workers are more likely to get training. The two samples are generally similar across observable dimensions, with notable differences being that individuals in our analysis are more likely to be full-time and hold a permanent contract.

In the next section, we present the results from estimating, by OLS and IV, a system of equations that resembles the model presented in section 3.2 for our application. This allows us to compare the results using our method to those obtained using standard approaches, the theoretical analysis of Subsection 3.2.2 having demonstrated the potential biases on OLS and IV estimators.

3.3.2 Preliminary analysis

We start by estimating the wage equation,

$$w_{it} = \alpha_t + \beta_t d_i + x_i \theta_t + v_{it}, \tag{3.5}$$

where w_{it} are log-wages at the end of 2013 ($t=1$), 2014 ($t=2$), and 2015 ($t=3$); d_i is an indicator for training between January 2014 and December 2015; and x_i are covariates that affect wages (as observed in 2013).[16] This equation is first estimated by OLS for each year separately, and then by 2SLS, instrumenting d_i by z_i, the *information on training* mentioned above. The estimations are done with and without controls. The DiD estimate of the effect of training in 2014 and 2015 is obtained as $\Delta\beta_2 = \beta_2 - \beta_1$ and $\Delta\beta_3 = \beta_3 - \beta_1$.

[16]For controls, we use: gender, age brackets, married, handicapped, having health problems, open-ended contract, full-time contract, socioeconomic status, firm size brackets, existence of an HR department, existence of wage incentives for performance (individual and collective), whether the firm outsources activities. See Table 3.3.1 for summary statistics.

Table 3.3.3: Static estimation of wage regressions with training

	OLS		2SLS	
	Without controls	With controls	Without controls	With controls
Log-wage levels				
2013	0.158	0.038	0.179	0.057
	(0.009)	(0.006)	(0.060)	(0.053)
2014	0.166	0.040	0.219	0.098
	(0.009)	(0.006)	(0.061)	(0.053)
2015	0.169	0.043	0.216	0.093
	(0.009)	(0.006)	(0.062)	(0.054)
Log-wage changes				
2014	0.007	0.002	0.040	0.041
	(0.003)	(0.003)	(0.019)	(0.025)
2015	0.011	0.005	0.037	0.036
	(0.003)	(0.004)	(0.022)	(0.030)
Nb of workers	7,533	7,533	7,533	7,533

The results are reported in Table 3.3.3. The OLS results suggest very small effects of training (differences in the β's around 0.2-0.5% with controls) and the effect of training on pre-treatment wages remains significant even after adding many controls to the estimation. After instrumenting the training variable, we see both stronger effects of around 4%, and the effect of training on 2013 wages stops being significant when controls are included in the regressions. Note that standard errors jump by one order of magnitude, pointing at a certain weakness of the instrument.

These results suggest the existence of a causal link between wages and training of around 4%, which is non negligible. We now use our model in order to check whether there is any reason to doubt that the IV estimation delivers an unbiased estimate of the causal effect of training on wages.

3.3.3 Parametric specification

In practice, we specify a parametric version of the model and we use maximum likelihood for estimation.

We assume that log-wages are normal conditional on type and training, and first-order autoregressive with autocorrelation coefficient ρ. More precisely, we postulate that

$$w_1 = \mu_1(h) + u_1, \quad \text{where} \quad u_1 \sim \mathcal{N}(0, \sigma_1^2(h)),$$

and for $t = 2,3$,

$$w_t = \mu_t(h,d) + u_t, \quad \text{where} \quad u_t \sim \mathcal{N}(\rho u_{t-1}, \sigma_t^2(h,d)).$$

Then, with $\varphi(u) = (2\pi)^{-1/2} e^{-u^2/2}$, we have,

$$f_1(w_1|h) = \frac{1}{\sigma_1(h)} \varphi\left(\frac{w_1 - \mu_1(h)}{\sigma_1(h)}\right),$$

and

$$f_{2|1}(w_2|w_1,h,d) = \frac{1}{\sigma_2(h,d)} \varphi\left(\frac{w_2 - \mu_2(h,d) - \rho[w_1 - \mu_1(h)]}{\sigma_2(h,d)}\right),$$

$$f_{3|2}(w_3|w_2,h,d) = \frac{1}{\sigma_3(h,d)} \varphi\left(\frac{w_3 - \mu_3(h,d) - \rho[w_2 - \mu_2(h,d)]}{\sigma_3(h,d)}\right).$$

The model is flexible at first and second order as parameters μ_t, σ_t are left unrestricted. A more flexible distribution than the normal could be used for the distribution of innovation errors, but, as we shall see, the link between wages and training is tiny. Thus, there is little data to infer higher order moments.

Probabilities $\pi(h,z,d)$ are left unrestricted.

The data for each individual i is the array $x_i = (w_{i1}, w_{i2}, w_{i3}, z_i, d_i)$. The parameters of the model are denoted $\beta = (\mu, \pi, \rho, \sigma)$. The complete likelihood of individual i's observations x_i and any type h is

$$\begin{aligned} \ell_{ih}(\beta) &\equiv \ell(x_i, h, \beta) \\ &= \pi(h, z_i, d_i)\, f_1(w_{i1}|h,\beta)\, f_{2|1}(w_{i2}|w_{1i}, h, d_i, \beta)\, f_{3|2}(w_{i3}|w_{i2}, h, d_i, \beta). \end{aligned} \tag{3.6}$$

The individual likelihood is $\ell_i(\beta) = \sum_h \ell_{ih}(\beta)$. The sample likelihood is the product of individual likelihoods, $L(\beta) = \prod_i \ell_i(\beta)$.

3.3.4 Types and likelihood maximization

We found that a two-stage approach to estimation worked best in our application. In the first stage, we classify workers into types based solely on their wages, abstracting from training and training information. We then perform a second round of classification within each group from the first stage, now allowing wages to depend on training. Within each stage, the EM algorithm is used to estimate the discrete mixture, and groups are labeled by increasing values of mean wages in 2013, i.e. by μ_1. Specifically, we now assume that each type h is a pair $h = (k,g)$, where $k \in \{1,\ldots,K\}$ is the first-stage type component (depending only on wages) and $g \in \{1,\ldots,G\}$ is a second-stage type component (depending

on training and wages). It follows that the total number of discrete groups is $H = GK$.[17]

We started by estimating the full model with unrestricted types in a single step. However, most of the latent classification was used to fit the overall distribution of wages, and little heterogeneity was spared to fit different relationships between wages and training. In particular, it was not possible to avoid wages in 2013 varying with training in 2014 within each estimated group. By first estimating a discrete mixture of wages and then re-estimating a discrete mixture of wages *and* training, given the first classification, we increase our chances of zooming in on the wage-training link. Note that, in principle, this two-stage procedure is used without loss of generality. Indeed, nothing prevents the estimated second-stage mixture from being exactly the same within each of the first-stage groups.

The simplified first-stage model is

$$\begin{aligned} w_1 &= \bar{\mu}_1(k) + u_1, \quad u_1 \sim N\left(0, \bar{\sigma}_1^2(k)\right), \\ w_t &= \bar{\mu}_t(k) + u_t, \quad u_t \sim N\left(\bar{\rho} u_{t-1}, \bar{\sigma}_t^2(k)\right), \quad t = 2, 3, \end{aligned}$$

where we use an upper bar to distinguish the first-stage variables and parameters from those of the second-stage.

We use a sequential EM-algorithm for the likelihood maximization of both stages (see Appendix B for details). Moreover, we relabel groups k and subgroups g by increasing values of $\overline{\mu}_1(k)$ and $\mu_1(k,g)$ after estimation has converged.

3.3.5 Bootstrap

Standard calculations of parameter standard errors do not incorporate the random nature of the estimated classification (even if it should be negligible asymptotically). We therefore bootstrap standard errors by resampling and reestimating many times the whole procedure. This is computationally intensive as we use 500 replicated samples, with replacement, from the original sample. Specifically, we use the weighted-likelihood bootstrap. O'Hagan et al. (2019) show that it provides a robust solution in our setting. Standard bootstrap may generate unstable results if re-sampling causes certain types to be under-represented or even to disappear. The weighted version draws non-zero weights for each observation from a Dirichlet distribution to ensure that no observations are completely dropped in any bootstrapped sample (Newton and Raftery, 1994). The weights λ_i are such that they sum to the size of the full sample, that is, $\sum_i \lambda_i = N$. We use the original, full-sample estimates as initial values for the algorithm at the beginning of each re-estimation. Confidence intervals can then be estimated by selecting the corresponding

[17] We could have let G depend on k, each first-stage type k determining a different number of second-stage types $G(k)$, but to keep the analysis relatively simple, we keep G constant across k.

percentiles of the bootstrapped parameter estimates, i.e. the 5th and 95th percentiles for a 90% confidence interval.

3.4 Results

3.4.1 Choosing the number of types (K and G)

Our estimation strategy requires the econometrician to choose the number of types in both stages, K and G.

Figure 3.4.1 presents some of the criteria we use to choose the number of types for the remainder of our analysis. In Figure 3.4.1(a), the different broken lines show how total likelihood ($\ln L$) and penalised-likelihood criteria evolve with K. The first two penalized-likelihood criteria are the well-known Akaike and Bayesian Information criteria (respectively, AIC and BIC). We are looking for "elbows", that is, values of K where the marginal gain in likelihood for an additional type is noticeably less than it is for $K-1$. There is a clear elbow at $K=3$ or $K=4$ for AIC and BIC. The third criterion, ICL (for *Integrated Conditional Likelihood*) is more or less steadily decreasing for all K. This criterion was proposed by Biernacki et al. (2000) to counter the tendency of BIC to overestimate the number of groups by penalizing the likelihood when groups are not well separated.[18]

In panel (b) of Figure 3.4.1 we study what happens to the sizes and means of the groups as we increase K. The groups all appear distinct when compared by their means, but very small groups start to appear for $K=6$ or 7. Notice also that there is only a small fraction of the workers who display different mean wages across periods. There is only one group with clearly different wage means for $K \leq 5$. For $K=6$ or 7 we see more than one group with different wage means, but this looks like a dilution of the only such group appearing when $K=4$ or 5. Combining the evidence from both panels of Figure 3.4.1, we choose $K=4$ for the first stage, and we leave to the second stage the task of determining the role of training in generating the observed changes on mean wages over time.

Figure 3.4.2 represents the results from the second stage of our estimation procedure. In panel (a), each of the four subpanels shows the likelihood criteria for each one of the four types obtained in the first stage. The likelihood, here, is a weighted sum of individual likelihoods where the weights are the posterior type probabilities estimated from the first stage (see Appendix B). For all types except $k=2$, the BIC wants $G=3$.[19] When $k=2$,

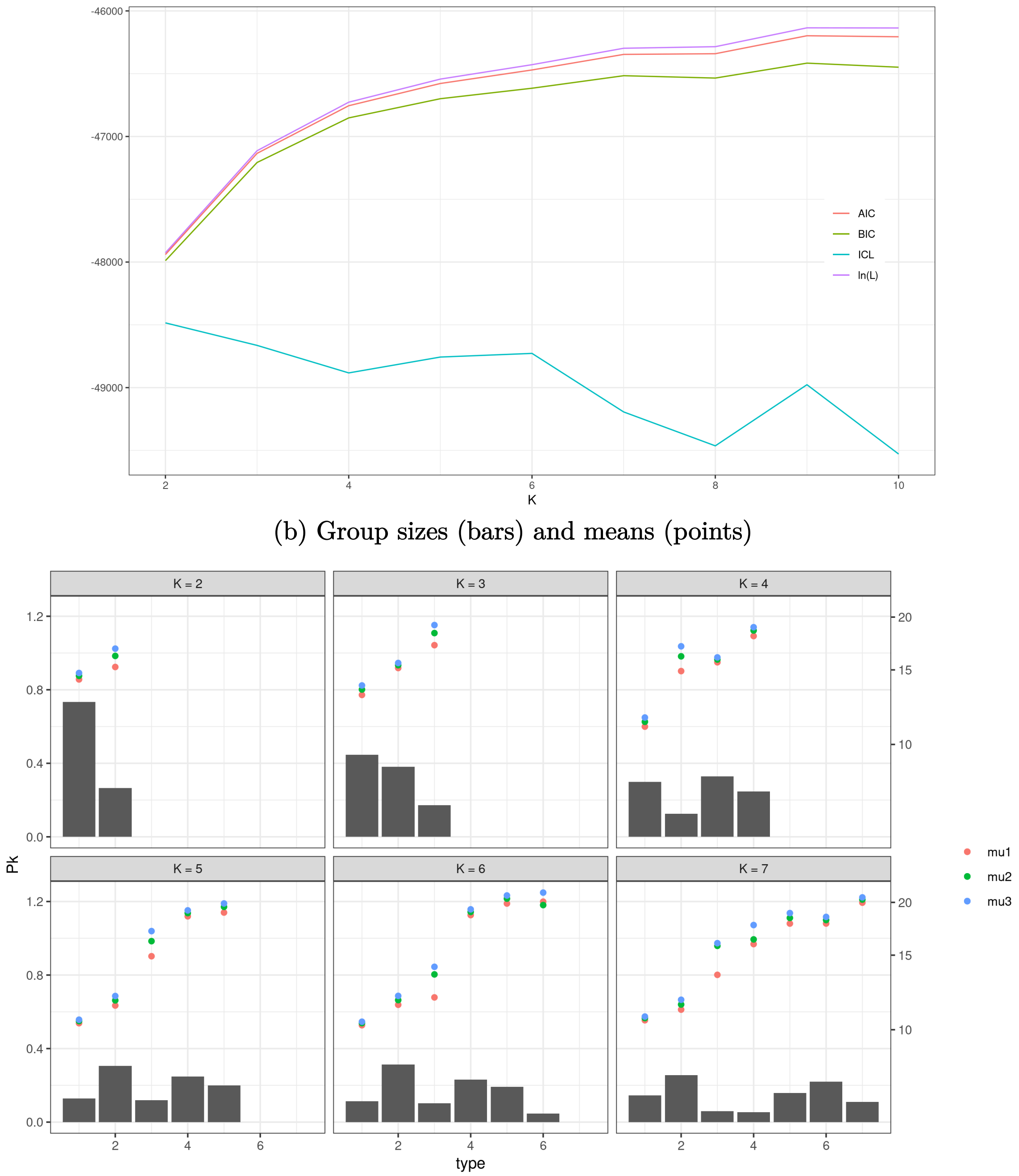
[18]As pointed out by Biernacki *et al.*, the BIC is a reliable approximation of the integrated likelihood if the estimated parameters are well within their domain. This is not the case if the estimated K is greater than the true one K^0, as $K-K^0$ shares should be equal to 0.

[19]The second-stage ICL actually wants a larger G than the BIC. Given the motivation for the ICL (the BIC can *overestimate* K) we choose the K suggested by the BIC in the second stage.

Figure 3.4.1: Choosing the number of types (K)

(a) Likelihood criteria

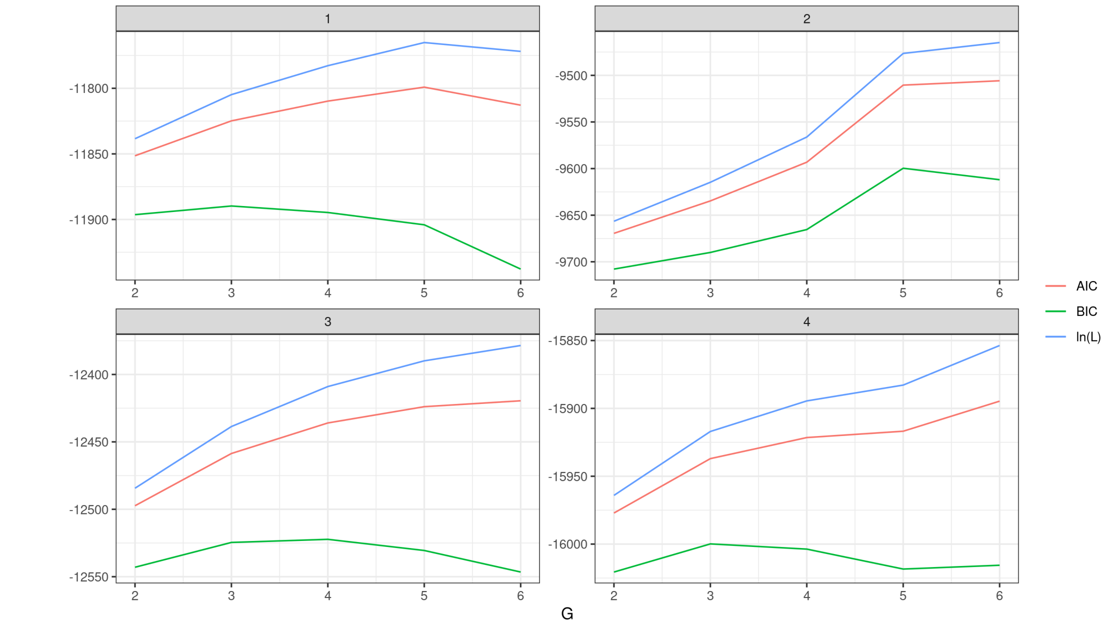
(b) Group sizes (bars) and means (points)



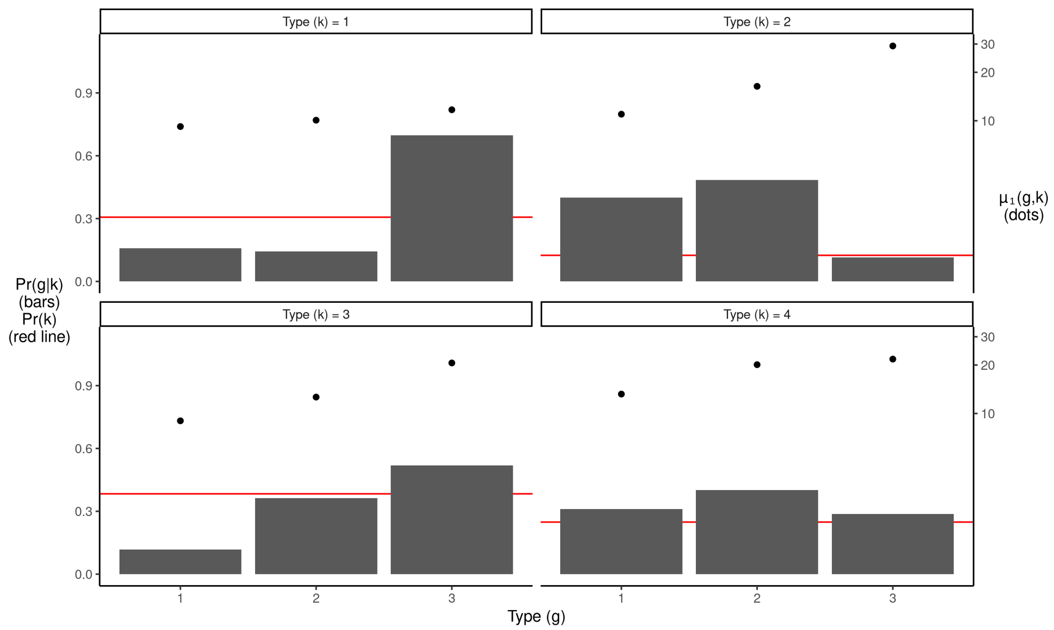
Notes: (a) If M is the number of parameters, N the number of observations, and L the likelihood, $AIC = -\ln L + \frac{1}{2}\ln M$, and $BIC = -\ln L + \ln(N)M$. We plot $-AIC$ and $-BIC$ on the figure. The ICL is an alternative criterion proposed by Biernacki et al. (2000). (b) The bars in Panel (b) are the shares of each group. The colored dots are the levels of estimated mean wages in the three years for which we observe wages.

Figure 3.4.2: Choosing G (stage 2)

(a) Likelihood criteria

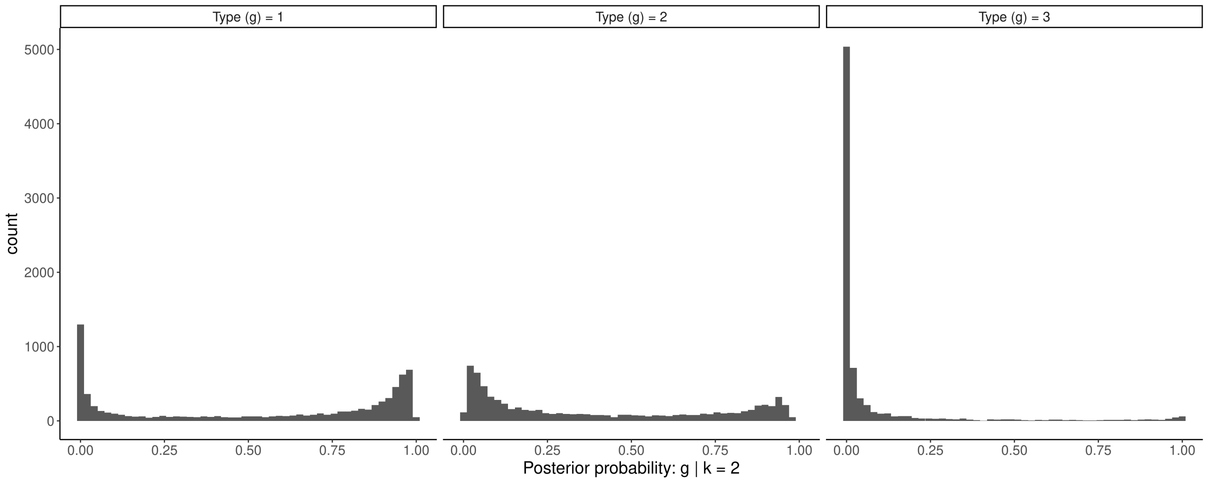


(b) Group sizes (bars) and means (points), $G=3$



Notes: (a) If M is the number of parameters, N the number of observations, and L the likelihood, $AIC = -\ln L + \frac{1}{2}\ln M$, and $BIC = -\ln L + \ln(N)M$. We plot $-AIC$ and $-BIC$ on the figure. (b) The bars in Panel (b) are the shares of each group. The dots are the estimated mean wages in 2013.

Figure 3.4.3: Posterior type probabilities $p_i(k=2,g)$



the optimal G is 5, but given that this is the smallest group from stage 1 (represented by the red horizontal lines in panel (b)), and for the sake of simplicity, we choose the same number G of second-stage types within each first-stage type k. To avoid an abundance of numbers and plots, we choose $G=3$ and show results with $G=3$ for all types $k=1,\ldots,4$. We checked that $G=5$ delivers similar conclusions.

Lastly, another criterion to determine the optimal number of groups is whether groups are well differentiated or not. This is what the ICL criterion takes into account, and the BIC does not. We can check the quality of the classification by plotting the distribution of posterior group-probabilities, $p_i(h)$.[20] Figure 3.4.3 displays these distributions for $k=2$ and all $g=1,2,3$. We see that they are concentrated near zero and one.

3.4.2 Observed characteristics by type

We did not include controls when estimating the model to avoid the double complication of choosing a functional form for probabilities $\pi(h,z,d)$ and specifying the interaction between observed and unobserved heterogeneity in these probabilities and the wage densities. But we can still study if types can be characterized by some specific values of observed variables. We first assign to each individual a type corresponding to their highest posterior probability, and then study the individuals assigned to each group. The results of this exercise are in Table 3.4.1. The employee characteristics are in panel (a), while the characteristics of the firms that employ individuals of that type are in panel (b).

Interestingly, although we only use wages in the first-stage, and wages and training in the second, the resulting classification does not correspond to any (obvious) classification in terms of other observed characteristics. For example, k is not obviously associated to education, and g is not obviously related to occupation or firm size.

[20]By definition, $p_i(h)=\Pr\{h_i=h\,|\,w_{it},z_i,d_i\}=\ell_{ih}(\widehat{\beta})/\sum_h \ell_{ih}(\widehat{\beta})$.

Table 3.4.1: Comparing types by baseline characteristics

	$k_i = 1$			$k_i = 2$			$k_i = 3$			$k_i = 4$		
$g_i =$	1	2	3	1	2	3	1	2	3	1	2	3
Demographics:												
Age (modal group)	45–49	45–49	40–44	50–54	40–44	55–59	40–44	50–54	50–54	45–49	45–49	50–54
Male	57.8	66.7	71.1	62.3	68.5	91.1	56.4	73.3	80.7	75.4	81.9	85.2
French	93.1	96.4	96.9	96.9	98.1	94.9	93.9	97.4	97.4	96.2	97.6	97.7
In couple	61.7	66.7	73.4	73.0	78.1	87.3	64.6	75.3	82.2	77.1	82.2	84.4
Has children	44.9	52.9	57.1	53.5	62.1	63.3	48.2	57.4	66.0	57.3	67.2	64.3
Disability	13.8	12.4	9.90	21.1	7.73	5.06	9.2	8.92	4.50	12.2	3.29	5.28
Previous health issue	3.29	5.33	3.86	6.92	2.40	1.27	6.78	4.14	1.69	4.06	1.90	1.51
Education:												
Less than HS diploma	59.6	63.6	46.5	54.7	22.4	13.9	61.5	50.5	18.1	47.3	20.3	21.4
HS diploma	24.9	19.1	22.4	16.0	16.0	8.86	23.0	19.1	13.6	17.9	15.8	12.3
Trade / voc. degree	9.88	10.7	18.3	12.9	22.4	15.2	7.99	17.9	26.0	17.4	23.9	20.9
Bachelor's degree	2.99	3.11	6.04	5.03	8.00	8.86	3.39	6.16	6.04	4.53	6.58	5.53
Master's degree +	2.10	2.22	5.93	10.10	29.9	53.2	2.66	5.98	35.7	12.4	32.9	39.2
Occupation:												
Unskilled blue collar	13.2	15.1	9.38	8.81	1.60	2.53	15.3	9.02	1.33	7.88	1.52	1.26
Skilled, technician	33.2	40.4	32.8	34.6	10.1	5.06	32.9	31.4	7.96	28.2	11.4	6.28
Office, public sector	43.4	27.6	28.8	26.7	10.9	5.06	42.9	25.7	6.71	23.2	8.99	7.79
Foreman/Supervisor	0.90	3.56	11.1	9.12	12.5	1.27	1.21	12.8	12.0	11.7	13.7	10.6
Technician, sales	2.69	6.67	9.55	6.29	10.1	3.80	2.66	12.0	9.36	7.88	10.1	4.02
Engineer, manager	0.60	2.22	5.64	9.43	53.3	81.0	0.73	7.18	61.5	19.1	53.4	68.8
Job characteristics:												
Log(hourly wage)												
2013 (w_1)	2.19	2.26	2.44	2.35	2.80	3.44	2.18	2.53	3.07	2.60	3.04	3.25
2014 (w_2)	2.21	2.28	2.47	2.45	2.89	3.50	2.19	2.54	3.09	2.66	3.06	3.27
2015 (w_3)	2.23	2.29	2.50	2.45	3.01	3.53	2.20	2.55	3.10	2.67	3.08	3.29
Permanent contract	97.6	98.2	97.8	96.2	98.7	96.2	98.5	99.2	99.3	97.9	98.6	98.2
Full time contract	87.7	88.9	94.9	93.4	97.9	94.9	86.7	95.5	96.7	95.9	96.8	97.5
Info. on training (z)	64.4	63.1	78.2	75.2	82.1	75.9	61.3	73.0	78.0	74.7	79.9	62.1
Training (d)	12.3	2.22	50.2	36.5	57.6	45.6	12.6	34.9	62.9	17.7	83.7	40.5
Firm characteristics:												
Number of employees												
3 to 49	47.6	48.9	34.0	31.8	23.7	29.1	49.2	37.1	17.0	31.0	17.8	25.6
5 to 249	25.1	28.0	26.3	22.6	19.7	19.0	24.9	19.4	20.6	22.9	19.0	23.4
25 to 499	7.49	6.22	7.13	10.1	10.9	12.7	5.81	8.28	9.36	11.7	9.37	11.6
50 to 999	4.19	3.11	6.50	8.49	8.80	3.80	6.05	8.28	9.06	8.35	8.23	9.05
1000 to 1999	2.69	4.89	5.70	7.86	7.47	3.80	3.63	5.89	7.44	4.30	7.85	8.54
More than 2000	12.9	8.89	20.4	19.2	29.3	31.6	10.4	21.1	36.6	21.7	37.7	21.9
Has HR department	76.9	75.6	85.6	85.5	92.3	88.6	71.7	82.8	91.5	87.1	93.7	90.2
Individual incentives	53.9	49.8	64.5	62.6	75.5	77.2	44.1	63.8	77.1	67.8	81.8	71.6
Collective incentives	57.5	57.3	72.9	69.2	79.5	78.5	53.5	72.9	85.0	76.1	86.2	76.9
Outsources activity	29.3	27.1	35.2	43.7	43.7	40.5	27.4	34.6	47.9	38.2	45.1	44.7
No. of observations	334	225	1740	318	375	79	413	1090	1360	419	790	398

3.4.3 Parameter estimates (with $K=4$ and $G=3$)

Figure 3.4.4 displays the probability of being treated conditional on first- and second-stage types, and the value of the instrument, i.e., $\pi(d=1|k,g,z)$. The error bars indicate bootstrapped, 90% confidence intervals. There are two key features to note. First, right-blue bars are higher than left-red ones. This is evidence of instrument monotonicity, which holds almost perfectly: those who receive information on training are more likely to train across all types (except $(k,g)=(2,2)$). Second, the bars are generally increasing in both k and g.

In Figure 3.4.5, we study the correlation between the types and the instrument. As already emphasized, IV is equal to LATE only if the instrument and the type are independent, which would require the red and blue bars to be equal for each combination of k and g, i.e., that $\Pr(k,g|z=1)=\Pr(k,g|z=0)$. This assumption seems violated here, and there seems to be a pattern to the differences in bars. First, subgroups $g=1,2$ show similar differences by z, opposite to $g=3$. Moreover, groups $k=1,3$ and $k=2,4$ show opposite differences by z, which could be significant as groups $k=2,4$ are the ones exhibiting the greatest mean wage variations over time.

3.4.4 Treatment effects (with $K=4$ and $G=3$)

We now move on to the main objects of our analysis, the treatment effects. Panel (a) of Figure 3.4.6 displays treatment effects $ATE(k,g)$, conditional on all types (k,g). Noting that the y-axes differ between cells, we see substantial heterogeneity in treatment effects across types. The effects estimated for $k=2$ are five times larger than for the rest of the k-types. However, none of these conditional ATEs is very precisely estimated.

Note that we calculate an empirical wage mean in 2013 for the trained and the untrained, and corresponding pseudo-ATEs, although we estimated mean wages in 2013, $\mu_1(k,g)$, as independent of d. We did this calculation to check this independence assumption *ex post*. Unfortunately, the 2013 pre-treatment effects are of similar size to (or even greater than) 2014 and 2015 post-treatment effects for some types. Together with the large bootstrap standard errors, this confirms that the conditional ATEs are essentially not interpretable.

Panel (b) of Figure 3.4.6 displays aggregate treatment effects conditional on first-stage types only, $ATE(k)$, obtained by summing over g conditional $ATE(k,g)$ weighted by $\pi(g|k)$. The empty black-outlined bars are the $ATT(k)$, obtained in the same way, but using weights $\pi(g|k,d=1)$. Recall, the first-stage classification orders workers by increasing abilities, a source of heterogeneity that determines wages independently of adult training. Generally the $ATE(k)$'s are small — around one percent or less — though for $k=2$ they are closer to three percent in 2014 and 2015, but imprecisely estimated. The $ATT(k)$'s are only marginally greater than the $ATE(k)$'s. This suggests workers are

Figure 3.4.4: Treatment probability, $\pi(d=1|k,g,z)$

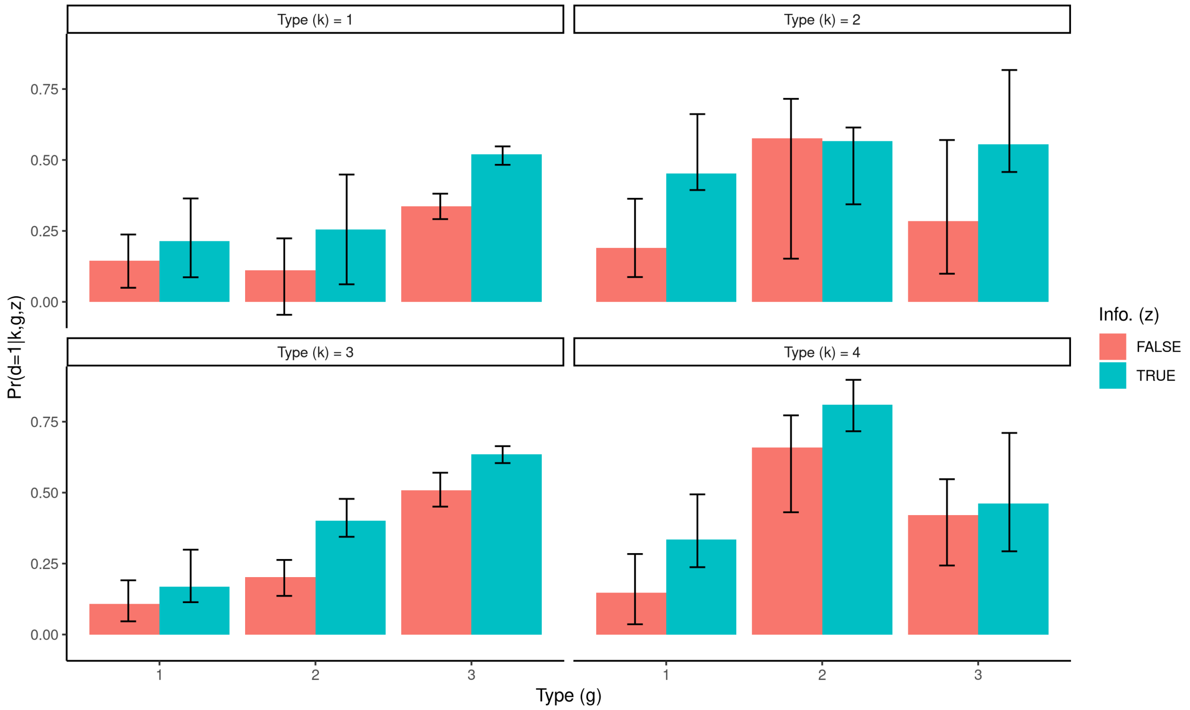


Figure 3.4.5: Composition, $\pi(k,g|z)$

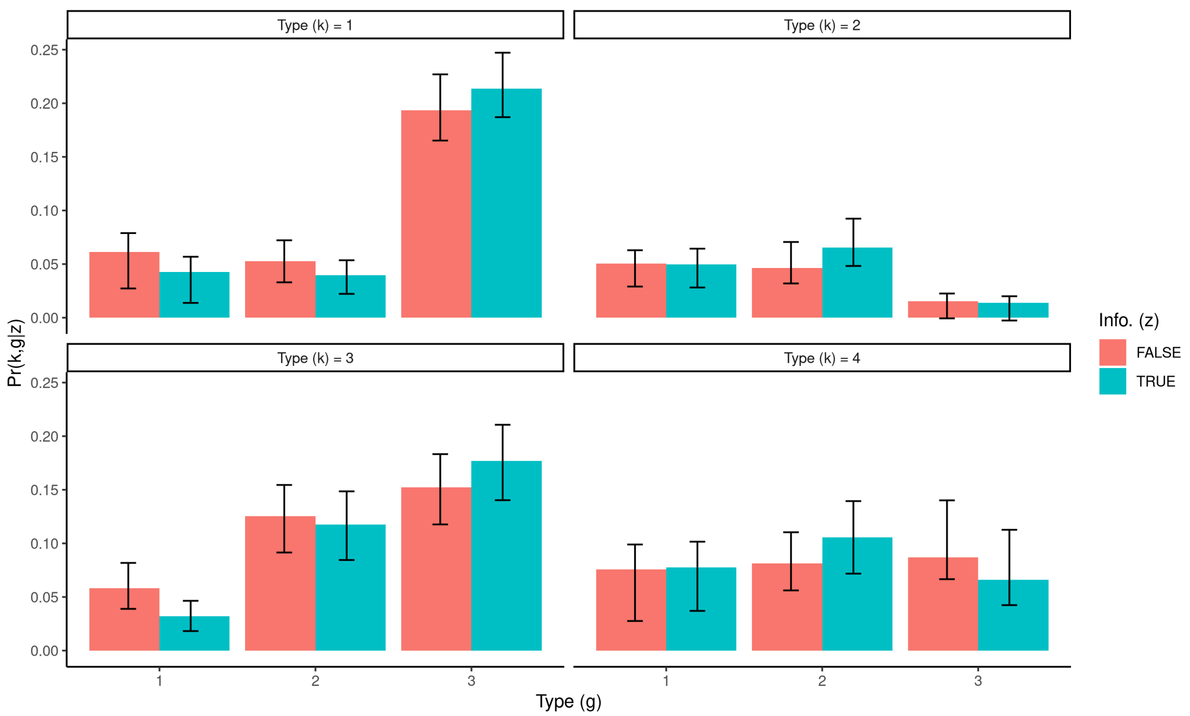
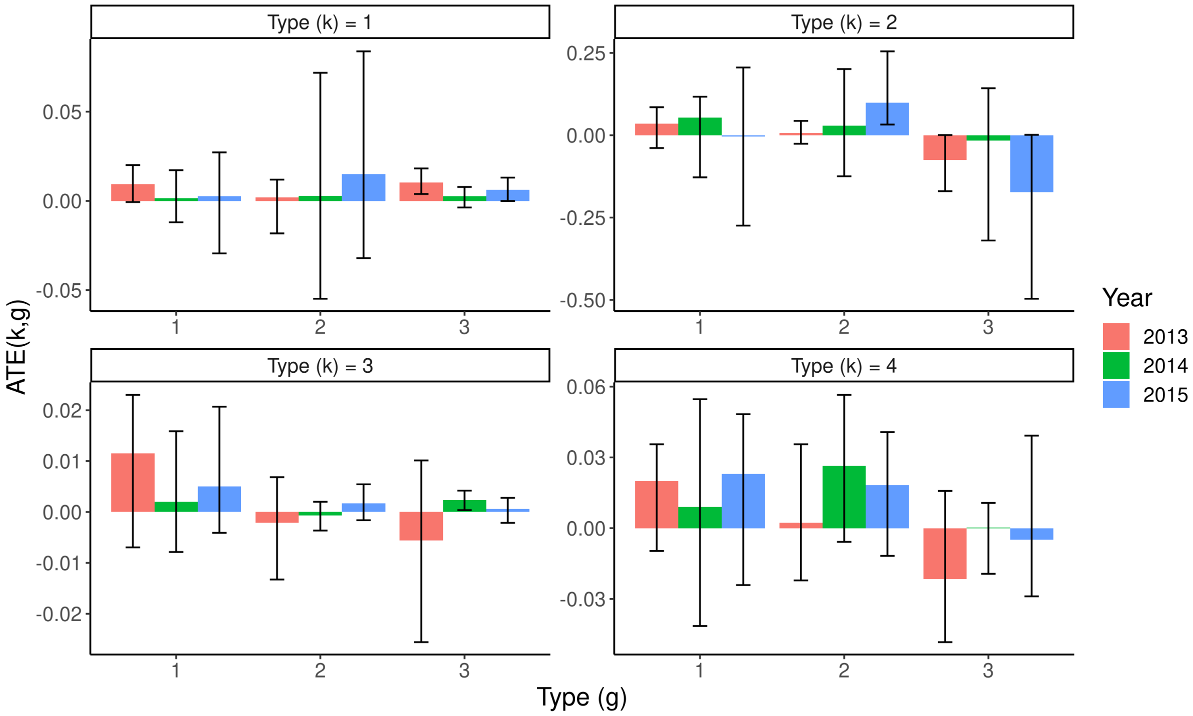
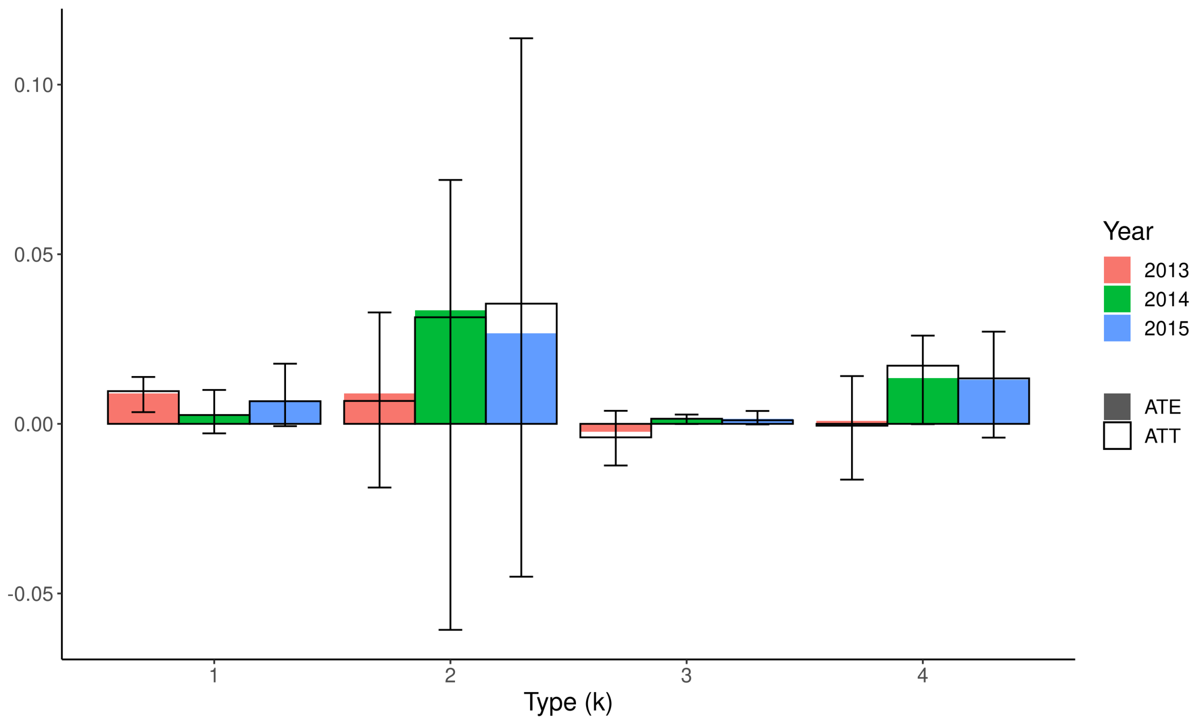


Figure 3.4.6: Type-conditional treatment effects

(a) $ATE(k,g)$



(b) $ATE(k)$ and $ATT(k)$



not selecting into training based on their *ex post* wage returns. Generally the picture in Figure 3.4.6 suggests small positive wage returns to training for most individuals of less than 1%, with a small number (around 12%) of individuals enjoying higher wage returns of about 3%. Lastly, note that the 2013 returns to training vanish after aggregating within the first-stage types (k).

Finally, we aggregate across both types k and g to obtain a variety of treatment effects summarising the whole sample, which are presented in Table 3.4.2. The first three columns are the average treatment effects weighted differently:

$$\begin{aligned}
ATE &= \sum_{h=(k,g)} \pi(h)\, ATE(h), \\
ATT &= \sum_{h=(k,g)} \pi(h|d=1)\, ATE(h), \\
LATE &= \sum_{h=(k,g)} \frac{\pi(h,d=1|z=1)-\pi(h,d=1|z=0)}{\sum_{h=(k,g)}\left[\pi(h,d=1|z=1)-\pi(h,d=1|z=0)\right]} ATE(h).
\end{aligned}$$

Then, in subsection 3.2.2 we established two potential biases on OLS and 2SLS estimators:

$$\begin{aligned}
b_{OLS} &= ATT + B_{OLS}, \\
b_{IV} &= LATE + B_{IV}.
\end{aligned}$$

Note that, in these formulas, b_{OLS} and b_{IV} refer to the population parameters under the model's null hypothesis. They are given by the formulas in Section 3.2.2.

However, while for OLS the above decomposition is exact, for IV there is an additional bias arising because, in the sample, pre-treatment wages may be correlated with treatment and instrument, and post-treatment wages may be correlated with the instrument given treatment. A similar situation occurs in standard 2SLS, with superfluous instruments leading to an overidentified model: the estimated residuals will not be exactly orthogonal to all the instruments but one. An important difference is that even in the just-identified case, the estimated residuals in our framework may still depend on the instrument; this cannot happen for 2SLS by construction.

The aggregate treatment effects, along with OLS and IV estimates and their associated biases, are displayed in Table 3.4.2. The top rows show the results obtained when the outcome is log-wage in levels. The bottom rows show results when the outcome is the difference in log-wages between pre- (2013) and post-treatment (2014 and 2015) periods.

We find similar sized estimates of ATE, ATT and LATE, of around 1%, with a big bootstrap standard error. The treatment effects calculated for wages in 2013 are much lower than those calculated for wages in 2014 and 2015, which is consistent with our identifying restriction.

Table 3.4.2: Aggregate treatment effects

	ATE	ATT	LATE	$\hat{b}_{OLS} = b_{OLS}$	B_{OLS}	$\hat{b}_{IV}$	b_{IV}	B_{IV}	$\hat{b}_{IV} - b_{IV}$
Log-wage levels									
2013	0.003	0.002	0.006	0.158	0.156	0.179	0.188	0.182	-0.011
	(0.004)	(0.004)	(0.005)	(0.027)	(0.029)	(0.063)	(0.060)	(0.061)	(0.021)
2014	0.009	0.010	0.012	0.164	0.153	0.219	0.199	0.187	0.022
	(0.006)	(0.008)	(0.008)	(0.027)	(0.035)	(0.063)	(0.060)	(0.061)	(0.023)
2015	0.009	0.011	0.010	0.167	0.157	0.216	0.204	0.194	0.010
	(0.007)	(0.006)	(0.009)	(0.027)	(0.029)	(0.063)	(0.060)	(0.061)	(0.024)
Log-wage changes									
'14 vs '13	0.009	0.010	0.012	0.008	-0.002	0.040	0.015	0.003	0.025
	(0.006)	(0.008)	(0.008)	(0.009)	(0.022)	(0.017)	(0.011)	(0.013)	(0.023)
'15 vs '13	0.009	0.011	0.010	0.012	0.001	0.037	0.020	0.010	0.017
	(0.007)	(0.006)	(0.009)	(0.008)	(0.024)	(0.021)	(0.011)	(0.009)	(0.024)

Notes: (1) The ATE, ATT and LATE estimates for 2013 log-wage levels are zero by Assumption 12. To obtain the (nonzero) values in the top rows of the table, we compute $\mu_1(h,d)$ as mean log-wages weighted by posterior probabilities separately for trained and untrained workers. For log-wage differences, the ATE, ATT and LATE refer to $\mu_t(h,d) - \mu_1(h)$. (2) Standard errors are in parentheses, calculated as the standard deviation of the parameter estimates from 500 weighted-likelihood bootstrap repetitions. (3) $\hat{b}_{OLS}$ and $\hat{b}_{IV}$ are "naive" estimates obtained using ordinary least squares (OLS) and two-stage least squares (IV). b_{OLS} and b_{IV} are our model analogues of these estimates, calculated using the formulas in subsection 3.2.2.

The biases resulting from heterogeneous treatment and counterfactual wage levels, B_{OLS} and B_{IV}, are of the same order of magnitude as the respective OLS and IV estimates, $\hat{b}_{OLS}$ and $\hat{b}_{IV}$.[21] This was expected: we already know that there is a lot of heterogeneity in wage trajectories. The statistical errors on the IV estimator $(\hat{b}_{IV} - b_{IV})$ — which arise because wages are not completely independent of the instrument — are small in comparison. Yet, they are similar in magnitude to treatment effects, but with a much larger bootstrap standard error.

In the bottom part of Table 3.4.2 we show the difference-in-difference (DiD) decomposition. Under the identifying restrictions that wages do not depend on the instrument given treatment and that pre-treatment wages do not depend on the treatment, DiD treatment effects and level treatment effects are identical; but the decompositions differ. We see that the bias due to heterogeneous trends — i.e. due to violation of the common trend assumption — is negligible for OLS (B_{OLS}) and is slightly bigger for IV (B_{IV}).[22] It is therefore likely that the sizeable IV estimate for wage differences (4%) is for a large part (maybe half of it) just noise.

All this evidence suggests that the effect of adult training on wages is very small, quasi undetectable.

[21] These biases are also much less precisely estimated than the treatment effects.
[22] This bias is zero by construction for OLS, hence its omission from Table 3.4.2.

3.5 Conclusion

In this article, we developed and demonstrated the empirical use of a novel methodology for estimating treatment effects. The method allows for unobserved heterogeneity. The identification of conditional treatment effects given latent types (ATE, ATT and LATE) is rendered possible by a combination of nonparametric difference-in-difference and instrumental-variable inference. Neither monotonicity nor common trends are required for identification. In addition, we allow outcome variables (wages) to be Markovian given treatment and latent type. Moreover, by assuming discrete types, we permit unobserved heterogeneity to condition observed outcomes, treatments and instruments in a very general way. For example, no form of linearity nor homoscedasticity is required in constrast with factor models. This also allows us to base the estimation of a flexible parametric form of the model on the EM algorithm. Our method is generally applicable to other policy evaluation problems.

In our application using recent French data on training and wages, we find that formal training seems to have a small positive effect on wages, around 1% on average, except for a small fraction of workers for whom we find treatment effect values of around 3%. This result helps understand why adult training is not more common and why the perceived impact of training by workers is low.

Bibliography

Abadie, A. (2005): "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies*, 72, 1–19.

Abadie, A., J. Angrist, and G. Imbens (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91–117.

Abbring, J. H. and G. J. v. d. Berg (2003): "The Nonparametric Identification of Treatment Effects in Duration Models," *Econometrica*, 71, 1491–1517.

Acemoglu, D. and J.-S. Pischke (1998): "Why Do Firms Train? Theory and Evidence," *Quarterly Journal of Economics*, 113, 79–119.

——— (1999): "The Structure of Wages and Investment in General Training," *Journal of Political Economy*, 107, 539–572.

Allman, E. S., C. Matias, and J. A. Rhodes (2009): "Identifiability of parameters in latent structure models with many observed variables," *Annals of Statistics*, 37, 3099–3132.

Ashenfelter, O. C. (1978): "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics*, 60, 47–57.

Athey, S. and G. W. Imbens (2006): "Identification and inference in nonlinear difference-in-differences models," *Econometrica*, 74, 431–497.

Attanasio, O., A. Kugler, and C. Meghir (2011): "Subsidizing Vocational Training for Disadvantaged Youth in Colombia: Evidence from a Randomized Trial," *American Economic Journal: Applied Economics*, 3, 188–220.

Ba, B. A., J. C. Ham, R. J. LaLonde, and X. Li (2017): "Estimating (Easily Interpreted) Dynamic Training Effects from Experimental Data," *Journal of Labor Economics*, 35, 149–200.

Ballot, G., F. Fakhfakh, and E. Taymaz (2006): "Who Benefits from Training and R&D, the Firm or the Workers?" *British Journal of Industrial Relations*, 44, 473–495.

Bartel, A. P. (1995): "Training, Wage Growth, and Job Performance: Evidence from a Company Database," *Journal of Labor Economics*, 13, 401–425, publisher: [University of Chicago Press, Society of Labor Economists, NORC at the University of Chicago].

Biernacki, C., G. Celeux, and G. Govaert (2000): "Assessing a mixture model for clustering with the integrated completed likelihood," *IEEE Transactions on Pattern*

Analysis and Machine Intelligence, 22, 719–725, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Blundell, R., L. Dearden, C. Meghir, and B. Sianesi (1999): "Human capital investment: the returns from education and training to the individual, the firm and the economy," *Fiscal Studies*, 20, 1–23.

Bonhomme, S., K. Jochmans, and J.-M. Robin (2016a): "Estimating multivariate latent-structure models," *Annals of Statistics*, 44, 540–563.

——— (2016b): "Non-parametric estimation of finite mixtures from repeated measurements," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78, 211–229.

——— (2017a): "Nonparametric estimation of non-exchangeable latent-variable models," *Journal of Econometrics*, 201, 237–248.

Bonhomme, S., T. Lamadon, and E. Manresa (2017b): "A Distributional Framework for Matched Employer Employee Data," mimeo, University of Chicago.

——— (2019): "A Distributional Framework for Matched Employer Employee Data," *Econometrica*, 87, 699–739.

Bonhomme, S. and U. Sauder (2011): "Recovering Distributions in Difference-in-Differences Models: A Comparison of Selective and Comprehensive Schooling," *Review of Economics and Statistics*, 93, 479–494.

Bonnal, L., D. Fougere, and A. Serandon (1997): "Evaluating the Impact of French Employment Policies on Individual Labour Market Histories," *Review of Economic Studies*, 64, 683–713.

Booth, A. L. (1993): "Private Sector Training and Graduate Earnings," *Review of Economics and Statistics*, 75, 164–170.

Brodaty, T., B. Crepon, and D. Fougere (2001): "Using Matching Estimators to Evaluate Alternative Youth Employment Programs : Evidence from France, 1986-1988," in *Econometric Evaluations of Labour Market Policies*, ed. by M. Lechner and F. Pfeiffer, Physica, Heidelberg, 2000-25, 85–124.

Caliendo, M., D. A. Cobb-Clark, H. Seitz, and A. Uhlendorff (2016): "Locus of Control and Investment in Training," SOEPpapers on Multidisciplinary Panel Data Research 890, DIW Berlin, The German Socio-Economic Panel (SOEP).

Callaway, B. and T. Li (2019): "Quantile treatment effects in difference in differences models with panel data," *Quantitative Economics*, 10, 1579–1618.

Callaway, B. and P. H. Sant'Anna (2020): "Difference-in-Differences with multiple time periods," *Journal of Econometrics*, S0304407620303948.

Card, D., J. Kluve, and A. Weber (2010): "Active Labour Market Policy Evaluations: A Meta-Analysis," *Economic Journal*, 120, 452–477.

——— (2018): "What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations," *Journal of the European Economic Association*, 16, 894–931.

Carneiro, P., J. J. Heckman, and E. Vytlacil (2010): "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin," *Econometrica*, 78, 377–394.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating Marginal Returns to Education," *American Economic Review*, 101, 2754–2781.

Carneiro, P. and S. Lee (2009): "Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality," *Journal of Econometrics*, 149, 191–208.

Chernozhukov, V. and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261.

Crepon, B., M. Ferracci, G. Jolivet, and G. J. van den Berg (2009): "Active Labor Market Policy Effects in a Dynamic Setting," *Journal of the European Economic Association*, 7, 595–605.

de Chaisemartin, C. and X. d'Haultfoeuille (2020): "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 110, 2964–2996.

Dearden, L., H. Reed, and J. V. Reenen (2006): "The Impact of Training on Productivity and Wages: Evidence from British Panel Data," *Oxford Bulletin of Economics and Statistics*, 68, 397–421.

D'Haultfoeuille, X., S. Hoderlein, and Y. Sasaki (2021): "Nonparametric Difference-in-Differences in Repeated Cross-Sections with Continuous Treatments," Papers 2104.14458, arXiv.org.

Fialho, P., G. Quintini, and M. Vandeweyer (2019): "Returns to different forms of job related training," Tech. Rep. 231, OECD Social, Employment and Migration Working Papers.

Freyaldenhoven, S., C. Hansen, and J. M. Shapiro (2019): "Pre-event Trends in the Panel Event-Study Design," *American Economic Review*, 109, 3307–3338.

Gerfin, M. and M. Lechner (2002): "A Microeconometric Evaluation of the Active Labour Market Policy in Switzerland," *Economic Journal*, 112, 854–893.

Goux, D. and E. Maurin (2000): "Returns to firm-provided training: evidence from French worker-firm matched data," *Labour Economics*, 7, 1–19.

Grip, A. D. and J. Sauermann (2012): "The Effects of Training on Own and Coworker Productivity: Evidence from a Field Experiment," *Economic Journal*, 122, 376–399.

Gritz, R. M. (1993): "The impact of training on the frequency and duration of employment," *Journal of Econometrics*, 57, 21–51.

Haelermans, C. and L. Borghans (2012): "Wage Effects of On-the-Job Training: A Meta-Analysis," *British Journal of Industrial Relations*, 50, 502–528.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098.

Heckman, J. J., H. Ichimura, and P. E. Todd (1997): "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, 64, 605–654.

Heckman, J. J., R. J. Lalonde, and J. A. Smith (1999): "The Economics and Econometrics of Active Labor Market Programs," in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card, Elsevier, vol. 3, 1865–2097.

Heckman, J. J. and R. J. Robb (1985): "Alternative methods for evaluating the impact of interventions : An overview," *Journal of Econometrics*, 30, 239–267.

Heckman, J. J. and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738.

Henry, M., Y. Kitamura, and B. Salanie (2014): "Partial identification of finite mixtures in econometric models," *Quantitative Economics*, 5, 123–144.

Hu, Y. (2015): "Microeconomic models with latent variables: Applications of measurement error models in empirical industrial organization and labor economics," Tech. rep., Cemmap, Working Papers, CWP03/15.

Hu, Y. and S. M. Schennach (2008): "Instrumental variable treatment of nonclassical measurement error models," *Econometrica*, 76, 195–216.

Hu, Y. H. and M. Shum (2012): "Nonparametric identification of dynamic models with unobserved state variables," *Journal of Econometrics*, 171, 32–44.

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

Kasahara, H. and K. Shimotsu (2009): "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices," *Econometrica*, 77, 135–175.

Kluve, J., H. Schneider, A. Uhlendorff, and Z. Zhao (2012): "Evaluating continuous training programmes by using the generalized propensity score," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 175, 587–617.

Konings, J. and S. Vanormelingen (2015): "The Impact of Training on Productivity and Wages: Firm-Level Evidence," *Review of Economics and Statistics*, 97, 485–497.

Krueger, A. and C. Rouse (1998): "The Effect of Workplace Education on Earnings, Turnover, and Job Performance," *Journal of Labor Economics*, 16, 61–94.

LaLonde, R. J. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76, 604–620.

Lee, D. S. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76, 1071–1102.

Leuven, E. and H. Oosterbeek (2008): "An alternative approach to estimate the wage returns to private-sector training," *Journal of Applied Econometrics*, 23, 423–434.

Li, F. and F. Li (2019): "Double-Robust Estimation in Difference-in-Differences with an Application to Traffic Safety Evaluation," *arXiv:1901.02152 [stat]*, arXiv: 1901.02152.

Lynch, L. M. (1992): "Private-Sector Training and the Earnings of Young Workers," *American Economic Review*, 82, 299–312.

McCall, B., J. Smith, and C. Wunsch (2016): "Government-Sponsored Vocational Education for Adults," in *Handbook of the Economics of Education*, Elsevier, vol. 5, 479–652.

Newton, M. A. and A. E. Raftery (1994): "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap," *Journal of the Royal Statistical Society: Series B (Methodological)*, 56, 3–26.

OECD (2021): *Career Guidance for Adults in a Changing World of Work.*

O'Hagan, A., T. B. Murphy, L. Scrucca, and I. C. Gormley (2019): "Investigation of parameter uncertainty in clustering using a Gaussian mixture model via jackknife, bootstrap and weighted likelihood bootstrap," *Computational Statistics*, 34, 1779–1813.

PARENT, D. (1999): "Wages and Mobility: The Impact of Employer-Provided Training," *Journal of Labor Economics*, 17, 298–317.

PISCHKE, J.-S. (2001): "Continuous training in Germany," *Journal of Population Economics*, 14, 523–548.

RIDDER, G. (1986): "An Event History Approach to the Evaluation of Training, Recruitment and Employment Programmes," *Journal of Applied Econometrics*, 1, 109–126.

RODRIGUEZ, J., F. SALTIEL, AND S. S. URZUA (2018): "Dynamic Treatment Effects of Job Training," NBER Working Papers 25408, National Bureau of Economic Research.

SANDVIK, J., R. SAOUMA, N. SEEGERT, AND C. T. STANTON (2021): "Treatment and Selection Effects of Formal Workplace Mentorship Programs," Working Paper 29148, National Bureau of Economic Research.

SANT'ANNA, P. H. C. AND J. ZHAO (2020): "Doubly robust difference-in-differences estimators," *Journal of Econometrics*, 219, 101–122.

SASAKI, Y. (2015): "Heterogeneity and selection in dynamic panel data," *Journal of Econometrics*, 188, 236–249.

SCHOENE, P. (2004): "Why is the Return to Training So High?" *Labour*, 18, 363–378.

SHIU, J.-L. AND Y. HU (2013): "Identification and estimation of nonlinear dynamic panel data models with unobserved covariates," *Journal of Econometrics*, 175, 116–131.

3.A Proof of the Identification Theorem

The identification proof has four steps.

Step 1: Identifying restrictions. Consider first the joint probability of treatment $d_i = d$, instrument $z_i = z$, and wages w_{i1} (before treatment) and w_{i2}, w_{i3} (after treatment). We now drop index i to lighten notation. Mixing over unobserved types, we have

$$p(z,d,w_1,w_2,w_3) = \sum_h \pi(h,z,d)\, f_1(w_1|h)\, f_{2|1}(w_2|w_1,h,d)\, f_{3|2}(w_3|w_2,h,d).$$

This joint probability can be rewritten as

$$p(z,d,w_1,w_2,w_3) = \sum_h \pi(h,z,d)\, f_2(w_2|h,d)\, f_{1|2}(w_1|w_2,h,d)\, f_{3|2}(w_3|w_2,h,d),$$

where $f_2(w_2|h,d) = \int f_1(w_1|h)\, f_{2|1}(w_2|w_1,h,d)\, \mathrm{d}w_1$ and

$$f_{1|2}(w_1|w_2,h,d) = \frac{f_1(w_1|h)\, f_{2|1}(w_2|w_1,h,d)}{f_2(w_2|h,d)}.$$

By Assumption 13, wage distributions are discrete. For simplicity, say that wages take N possible values in all periods. Then, for each possible value of (z,d,w_2), we can store these probabilities $p(.)$ in a matrix

$$P(z,d,w_2) = [p(z,d,w_1,w_2,w_3)]_{w_1 \times w_3},$$

where the subscript $w_1 \times w_3$ means that the values of w_1 index rows and those of w_3 index columns. Let $F_1(d,w_2) = \left[f_{1|2}(w_1|w_2,h,d)\right]_{w_1 \times h}$ be the matrix of pre-treatment wage probabilities, with w_1 indexing rows and h indexing columns. Similarly, let $F_2(d,w_2) = \left[f_{3|2}(w_3|w_2,h,d)\right]_{w_3 \times h}$ be the post-treatment matrix. Finally, let

$$D(z,d,w_2) = \mathrm{diag}\left[\pi(h,z,d)\, f_2(w_2|h,d)\right]_h$$

be the diagonal matrix with $\pi(h,z,d)\, f_2(w_2|h,d)$ in the hth diagonal entry. In matrix notation, we then have, for every (w_2,z,d),

$$P(z,d,w_2) = F_1(d,w_2)\, D(z,d,w_2)\, F_2(d,w_2)^\top.$$

Step 2: Identification given treatment d and first post-treatment wage w_2. We first proceed to show that the matrices $F_1(d,w_2)$, $D(z,d,w_2)$ and $F_2(d,w_2)$ are identified (for each (w_2,d,z)). Importantly, $F_1(d,w_2)$ and $F_2(d,w_2)$ are independent of z as wages are independent of the instrument given treatment and type (Assumption 7). So, there are two observable matrices, $P(0,d,w_2)$ and $P(1,d,w_2)$, with the same algebraic structure.[23] Under Assumption 9, $F_1(d,w_2)$ and $F_2(d,w_2)$ are full-column rank, and under Assumption 8 the matrix $D(0,d,w_2)$ is invertible for all $w_2 \in \mathcal{W}_2(d)$ (i.e., the common support of all distributions $f_2(w_2|h,d)$, $h = 1,...,H$).

To simplify the notation, let us omit the dependence on (d,w_2). Matrix $P(0,d,w_2) := P(0)$ has rank H and there exists a singular value decomposition: $P(0) = U\Lambda V^\top$, where U and V are nonsingular (N,N) matrices with $U^\top U = I_N$, $V^\top V = I_N$ and Λ is a diagonal matrix of dimension N. The number of non-zero diagonal entries in Λ is equal to the number of groups H. Let Λ_1 be the (H,H) diagonal matrix containing the non-zero singular values, and let $U = (U_1,U_2)$ and $V = (V_1,V_2)$ partition the columns of Λ accordingly, so that $P(0) = U_1\Lambda_1 V_1^\top$.

[23]That is, the same two matrices F_1 and $F_2^\top$ are, respectively, pre-multiplying and post-multiplying a diagonal matrix D, that varies with z, for each (w_2,d).

Note also that since the columns of U are orthogonal vectors,

$$U_2^\top P(0) = U_2^\top U_1 \Lambda_1 V_1^\top = 0_{(N-H)\times N}.$$

Hence,

$$U_2^\top P(0) = U_2^\top F_1 D(0) F_2^\top = 0_{(N-H)\times N},$$

where $F_1(d, w_2) := F_1$ and $F_2(d, w_2) := F_2$. As $D(0)F_2^\top$ is a full row-rank (H, N) matrix, it follows that $U_2^\top F_1 = 0_{(N-H)\times H}$. A similar argument implies that $P(0)V_2 = 0$ since $V_1^\top V_2 = 0$. Now, since $F_1 D(0)$ has rank H, it follows that $F_2^\top V_2 = 0_{H\times(N-H)}$.

Next, using the singular value decomposition of $P(0)$, we have

$$\Lambda^{-1} U_1^\top F_1 D(0) F_2^\top V_1 = \Lambda^{-1} U_1^\top P(0) V_1 = V_1^\top V_1 = I_H$$

Given this result, if we define $W = \Lambda_1^{-1} U_1^\top F_1$, then, $W^{-1} = D(0) F_2^\top V_1$.

Now, similarly denoting $P(1, d, w_2) := P(1)$, we also find that

$$\Lambda_1^{-1} U_1^\top P(1) V_1 = \Lambda_1^{-1} U_1^\top F_1 D(1) F_2^\top V_1 = W D(1) D(0)^{-1} W^{-1}.$$

The diagonal entries of $D(1)D(0)^{-1}$ being distinct by Assumption 10, they are uniquely determined as the eigenvalues of the matrix $\Lambda_1^{-1} U_1^\top P(1) V_1$, which is derived from $P(1)$ and $P(0)$ only. However, eigenvectors are determined only up to a multiplicative constant. So, let $\widehat{W}$ be one matrix of eigenvectors. There exists a diagonal matrix Δ such that $\widehat{W} = W\Delta = \Lambda_1^{-1} U_1^\top F_1 \Delta$. Then, $\Lambda_1 \widehat{W} = U_1^\top F_1 \Delta$. As $U_2^\top F_1 \Delta = 0_{(N-H)\times H}$, we also have

$$\begin{pmatrix} \Lambda_1 \widehat{W} \\ 0_{(N-H)\times H} \end{pmatrix} = U^\top F_1 \Delta.$$

Hence,

$$U_1 \Lambda_1 \widehat{W} = U \begin{pmatrix} \Lambda_1 \widehat{W} \\ 0_{(N-H)\times H} \end{pmatrix} = U U^\top F_1 \Delta = F_1 \Delta.$$

That is, $F_1 \Delta = U_1 \Lambda_1 \widehat{W}$ is identified. We also have $F_1 = U_1 \Lambda_1 \widehat{W} \Delta^{-1}$. Since the rows of F_1 sum to one (each column being a probability distribution), then, it is easy to check that

$$(\Delta_1, \ldots, \Delta_H) = (1, 1, \ldots, 1) U_1 \Lambda_1 \widehat{W},$$

and hence, Δ is identified (since all its diagonal terms Δ_k are identified). If Δ is known, then F_1 is known.

Lastly, we have $\Delta \widehat{W}^{-1} = W^{-1} = D(0) F_2^\top V_1$. Applying the same argument as above,

we have that

$$\begin{aligned}
W^{-1}V_1^\top &= \left(D(0)F_2^\top V_1, 0_{H\times(N-H)}\right)\begin{pmatrix} V_1^\top \\ V_2^\top \end{pmatrix} \\
&= \left(D(0)F_2^\top V_1,\ D(0)F_2^\top V_2\right)V^\top \\
&= D(0)F_2^\top V V^\top \\
&= D(0)F_2^\top .
\end{aligned}$$

In the same fashion as above, the rows of F_2 summing to one, it follows that $D(0)$ and F_2 are identified. Hence $D(1)$ is also identified.

Step 3: Common labeling given d. In the previous step, we have estimated

$$D(1,d,w_2)D(0,d,w_2)^{-1} = \mathrm{diag}\left[\frac{\pi(h,1,d)}{\pi(h,0,d)}\right]_h .$$

By Assumption 10, these eigenvalues are all different (and independent of w_2). One can thus relabel groups for each d so that the labelling is consistent for all possible choices of w_2. By Assumption 11, Step 2 can be done for all wages w_2 in the common support, which is also the entire support. Thus, we can sum $D(0,d,w_2)$ and $D(1,d,w_2)$ over w_2 and eliminate $f_2(w_2|h,d)$ (which sums to one on its support). This identifies $\pi(0,h,d)$ and $\pi(1,h,d)$ for all h. Knowing the $\pi(h,z,d)$'s and $D(z,d,w_2)$, we identify $f_2(w_2|h,d)$. Since $f_{1|2}(w_1|w_2,h,d)$ is already identified, then, using Bayes' formula, $f_1(w_1|h)\, f_{2|1}(w_2|w_1,h,d)$ is also identified. Summing this function over w_2 then identifies $f_1(w_1|h)$. The conditional distribution $f_{2|1}(w_2|w_1,h,d)$ follows. The conditional distribution $f_{3|2}(w_3|w_2,h,d)$ is identified since it is given by F_2, which is identified.

Step 4: Common labeling across treatments. It remains to align the groupings across treatments. This is easily done by remarking that $f_1(w_1|h)$ is independent of d (Assumption 12) and therefore, can be used to make sure that the same groups have identical labels across treatments.
Q.E.D.

3.B Sequential EM-algorithm formulas

3.B.1 Stage 1

In stage 1 wages (and types) are independent of training, d, and information z. We denote first-stage types by k, and the number of types is K. We assume that log-wages are normal and denote log-wage in period t by w_t. The difference with the model described

above is that μ and σ depend only on the first-stage type k, and do not depend on d. To avoid ambiguity, we use an upper bar to distinguish the first-stage from the second-stage variables and parameters,

$$\begin{aligned} w_1 &= \bar{\mu}_1(k) + u_1, \quad u_1 \sim N\left(0, \bar{\sigma}_1^2(k)\right), \\ w_t &= \bar{\mu}_t(k) + u_t, \quad u_t \sim N\left(\bar{\rho} u_{t-1}, \bar{\sigma}_t^2(k)\right), \quad t = 2, 3. \end{aligned}$$

E-step. The complete individual likelihood in stage 1 has a simplified form (depending only on k), that is,

$$\bar{\ell}_{ik}(\beta) = \bar{\pi}(k) \bar{f}_1(w_{i1}|k) \bar{f}_{2|1}(w_{i2}|w_{i1}, k) \bar{f}_{3|2}(w_{i3}|w_{i2}, k) \tag{3.7}$$

The posterior probability of worker i to be of type k given data (i.e., the conditional probability of k knowing i, also called *responsibility*), denoted $\bar{p}_{ik}$, can be computed with the help of contributions to likelihood, using Bayes' rule. Let $\beta^{(m)}$ denote an estimate of the parameters at the end of iteration m. More precisely, we have,

$$\bar{p}_{ik}^{(m)} \equiv \frac{\bar{\ell}_{ik}(\beta^{(m)})}{\sum_k \bar{\ell}_{ik}(\beta^{(m)})}. \tag{3.8}$$

M-step. We update the parameters sequentially as follows, using the following sequential procedure.

1. Update pre-treatment wage distribution parameters $\overline{\mu}, \overline{\sigma}^2$ given current iteration $\bar{\rho}^{(m-1)}$ of the AR parameter as

$$\bar{\mu}_t^{(m)}(k) = \frac{\sum_i \bar{p}_{ik}^{(m)} w_{it}}{\sum_i \bar{p}_{ik}^{(m)}},$$

and, with $\bar{u}_{itk}^{(m)} = w_{it} - \bar{\mu}_t^{(m)}(k)$ for $t = 1$,

$$(\bar{\sigma}_1^2)^{(m)}(k) = \frac{\sum_i \bar{p}_{ik}^{(m)} \left(\bar{u}_{i1k}^{(m)}\right)^2}{\sum_i \bar{p}_{ik}^{(m)}},$$

and for $t = 2, 3$,

$$(\bar{\sigma}_t^2)^{(m)}(k) = \frac{\sum_i \bar{p}_{ik}^{(m)} \left[\bar{u}_{itk}^{(m)}(k) - \bar{\rho}^{(m-1)} \bar{u}_{i,t-1,k}^{(m)}\right]^2}{\sum_i \bar{p}_{ik}^{(m-1)}}.$$

2. Then update $\bar{\rho}$ as follows,

$$\bar{\rho}^{(m)} = \frac{\sum_i \sum_k \bar{p}_{ik}^{(m)} \left(\frac{\bar{u}_{i1k}^{(m)} \bar{u}_{i2k}^{(m)}}{(\bar{\sigma}_2^2)^{(m)}(k)} + \frac{\bar{u}_{i2k}^{(m)} \bar{u}_{i3k}^{(m)}}{(\bar{\sigma}_3^2)^{(m)}(k)} \right)}{\sum_i \sum_k \bar{p}_{ik}^{(m)} \left(\frac{(\bar{u}_{i1k}^{(m)})^2}{(\bar{\sigma}_2^2)^{(m)}(k)} + \frac{(\bar{u}_{i2k}^{(m)})^2}{(\bar{\sigma}_3^2)^{(m)}(k)} \right)}.$$

The standard EM procedure would have all $\overline{\mu}, \overline{\sigma}^2$ and $\overline{\rho}$ estimated by weighted nonlinear least squares. By simplifying the M-estimation in this way, we do not obtain efficient M-step updates but the sequential EM algorithm keeps increasing the likelihood at each iteration.

3. Finally, we update $\bar{\pi}$ as the mean posterior probability across all workers,

$$\bar{\pi}^{(m)}(k) = \frac{1}{N} \sum_i \bar{p}_{ik}^{(m)}.$$

We continue iterating between these steps until the algorithm converges. The key results from the first stage are the final posterior probabilities, $\bar{p}_{ik}$, which we use as weights throughout the second stage to "allocate" workers to types.

3.B.2 Stage 2

To understand our 2-stages procedure, it can be helpful to consider the hypothetical case of a perfect (or hard) classification in stage 1. If each individual i belonged to only one group (or type) with probability 1, then, we would run stage 2 on each type k, only including those individuals classified as that type. With our soft classification, the posteriors $\bar{p}_{ik}$ can take any value between zero and one. Therefore, each observation i can contribute to the estimation of the model of several types in stage 2.

We assume now that the distribution of log-wages in period t, still denoted w_t, now depend on (k,g) and treatment d. Wages are given by the following expressions,

$$w_1 = \mu_1(k,g) + u_1, \quad u_1 \sim N\left(0, \sigma_1^2(k,g)\right) \tag{3.9}$$

$$w_t = \mu_t(k,g,d) + u_t, \quad u_t \sim N\left(\rho u_{t-1}, \sigma_t^2(k,g,d)\right), \quad t = 2,3 \tag{3.10}$$

The complete individual likelihood for stage 2 is now given by:

$$\ell_{ikg}(\beta) = \pi(g,k,z_i,d_i)\, f_1(w_{i1}|k,g)\, f_{2|1}(w_{i2}|w_{i1},k,g,d_i)\, f_{3|2}(w_{i3}|w_{i2},k,g,d_i) \tag{3.11}$$

In stage 2, we run the following procedure for each type $k \in K$ obtained in stage 1. As in stage 1, we iterate between an E-step (in which we update the posterior probabilities) and

an M-step (in which we maximize the likelihood given the posteriors from the E-step). But we use the expression of p_{ik}, obtained in the first stage, to compute the posterior probabilities of all (k,g)s.

E-step. In the E-step, at the m-th iteration, we update the posterior probabilities as follows,

$$p_{ig|k}^{(m)} = \frac{\ell_{ikg}(\beta^{(m)})}{\sum_g \ell_{ikg}(\beta^{(m)})}. \tag{3.12}$$

Let also $p_{ikg}^{(m)} = \bar{p}_i(k) p_{ig|k}^{(m)}$ (using the estimated posterior probabilities $\bar{p}_i(k)$ from the first stage).

M-step. In the M-step we update the parameters of the likelihood function sequentially.

1. For $t = 1$:

$$\mu_1^{(m)}(k,g) = \frac{\sum_i p_{ikg}^{(m)} w_{i1}}{\sum_i p_{ikg}^{(m)}}, \tag{3.13}$$

$$(\sigma_1^2)^{(m)}(k,g) = \frac{\sum_i p_{ikg}^{(m)} (u_{i1kg}^{(m)})^2}{\sum_i p_{ikg}^{(m)}}, \tag{3.14}$$

   with $u_{i1kg}^{(m)} = w_{i1} - \mu_1^{(m)}(k,g)$.

2. Then, for $t = 2,3$,

$$\mu_t^{(m)}(k,g,d) = \frac{\sum_{\{i:d_i=d\}} p_{ikg}^{(m)} \left[w_{it} - \rho^{(m-1)} u_{i,t-1,kgd}^{(m)} \right]}{\sum_{\{i:d_i=d\}} p_{ikg}^{(m)}}$$

$$(\sigma_t^2)^{(m)}(k,g,d) = \frac{\sum_{\{i:d_i=d\}} p_{ikg}^{(m)} \left[u_{itkgd}^{(m)} - \rho^{(m-1)} u_{i,t-1,kgd}^{(m)} \right]^2}{\sum_{\{i:d_i=d\}} p_{ikg}^{(m)}},$$

   where $u_{itkgd}^{(m)} = w_{it} - \mu_t^{(m)}(k,g,d), t = 2,3$.
   Note that $\mu_t(k,g,d)$ now depends on ρ for $t = 2,3$ because we impose $\mu_1(k,g,0) = \mu_1(k,g,1) = \mu_1(k,g)$, i.e., treatment d has no effect on pre-treatment wages, conditional on type (k,g). If we relaxed this constraint, the estimator $\mu_t(k,g,d)$ would always be a simple weighted average of w_{it}.

3. Denote $I(d) = \{i : d_i = d\}$, then, we can update the autoregressive parameter ρ as

follows,

$$\rho^{(m)} = \frac{\sum_{k,g} \sum_{d \in \{0,1\}} \sum_{i \in I(d)} p_{ikg}^{(m)} \left(\frac{u_{i1kg}^{(m)} u_{i2kgd}^{(m)}}{(\sigma_2^2)^{(m)}(k,g,d)} + \frac{u_{i2kgd}^{(m)} u_{i3kgd}^{(m)}}{(\sigma_3^2)^{(m)}(k,g,d)} \right)}{\sum_{k,g} \sum_{d \in \{0,1\}} \sum_{i \in I(d)} p_{ikg}^{(m)} \left(\frac{(u_{i1kg}^{(m)})^2}{(\sigma_2^2)^{(m)}(k,g,d)} + \frac{(u_{i2kgd}^{(m)})^2}{(\sigma_3^2)^{(m)}(k,g,d)} \right)}.$$

4. Finally, the type-state probabilities $\pi(k,g,z,d)$ are estimated as the average of posterior probabilities

$$\pi^{(m)}(k,g,z,d) = \frac{1}{N} \sum_{\{i : z_i = z, d_i = d\}} p_{ikg}^{(m)}.$$

Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed sollicitudin massa vel venenatis dictum. Aliquam erat volutpat. Phasellus accumsan eu felis at luctus. Integer neque elit, venenatis sed iaculis in, tincidunt nec augue. Aliquam erat volutpat. Nulla sodales tortor non justo tincidunt, non varius risus mollis. Aliquam est purus, cursus at nulla ac, sollicitudin placerat diam. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Ut at leo eget metus scelerisque venenatis. Sed quis dui nisi. Morbi sodales, leo ac scelerisque malesuada, libero sem placerat ante, sit amet ullamcorper ligula nulla vestibulum tellus.

Institut d'études politiques de Paris
ÉCOLE DOCTORALE DE SCIENCES PO
Programme doctoral d économie
Département d'économie
Doctorat en sciences économiques

Essais sur les compétences et l'éducation

Oliver CASSAGNEAU-FRANCIS

Thèse dirigée par
Ghazala AZMAT, Professor of Economics, IEP de Paris
Soutenue le 7 juin, 2022

Jury:
Emla FITZSIMMONS, Professor of Economics, UCL (rapporteur)
Eric FRENCH, Professor of Economics, University of Cambridge (rapporteur)
Clement DE CHAISEMARTIN, Professeur des Universités, IEP de Paris
Robert GARY-BOBO, Professeur des Universités, Université Paris 1
Jean-Marc ROBIN, Professeur des Universités, IEP de Paris

Résumé

Les trois chapitres de cette thèse étudient différents aspects des compétences et du capital humain dans des contextes différents : les déterminants (chapitre 1) et les rendements (chapitre 2) de l'enseignement supérieur en Angleterre, et les rendements de la formation formelle en France (chapitre 3).

Le premier chapitre examine la question suivante : qu'est-ce qui pousse certains jeunes à poursuivre leur formation à l'université alors que d'autres ne le font pas ? La nouveauté de l'approche présentée dans ce chapitre est l'accent mis sur les facteurs non pécuniaires : à l'aide de données tirées d'enquêtes détaillées provenant d'études de cohortes britanniques, je compare les attentes des jeunes, qu'ils fréquentent ou non l'université, sur de nombreux aspects de la vie à l'université et après celle-ci. Il s'avère que les facteurs non pécuniaires, et en particulier non financiers, sont beaucoup plus importants que les salaires pour déterminer les choix des jeunes en matière d'enseignement supérieur.

Dans le deuxième chapitre, l'analyse se concentre sur les rendements espérés avant l'entrée à l'université, puis sur les rendements réalisés après l'université. En utilisant les mêmes études de cohortes que dans le chapitre 1, j'étudie le rendement salarial d'un diplôme universitaire au Royaume-Uni en fonction des capacités des étudiants à l'entrée à l'université. Je développe une méthodologie qui permet d'estimer les capacités préalables cognitives et non cognitives. J'utilise ces estimations pour étudier comment le rendement de l'université varie selon les groupes d'individus ayant des niveaux d'aptitudes différents, en tenant compte des interactions entre les différentes composantes de l'aptitude.

Dans le troisième chapitre, l'accent est mis sur la formation formelle plutôt que sur l'enseignement supérieur, toujours dans le cadre du thème des compétences et de l'investissement en capital humain. En utilisant une nouvelle méthodologie dans l'esprit de la méthode des doubles différences, qui permet spécifiquement l'hétérogénéité non observée. Nous exploitons les nouvelles données françaises sur la formation pour estimer les rendements salariaux de la formation formelle. Nous trouvons de faibles estimations des rendements de la formation de l'ordre de 1 à 3 %, ce qui suggère que les estimations plus importantes trouvées dans des études antérieures n'ont sans doute pas réussi à tenir pleinement compte de l'hétérogénéité non mesurée. Cette thèse souligne l'importance de l'enseignement supérieur dans l'explication des inégalités salariales, tout en mettant en évidence les multiples dimensions que les jeunes considèrent lorsqu'ils décident de faire

cet investissement pour leur avenir.

Chapitre 1 — Le rôle des revenus et des autres facteurs dans la fréquentation universitaire

Les sciences économiques se sont concentrées jusqu'à présent sur l'importance du bénéfice salarial des diplômés comme principal moteur de la fréquentation universitaire, et sur les contraintes de crédit comme principal obstacle à l'investissement dans le capital humain. Cependant, les chercheurs ont récemment mis en évidence l'incapacité de ces facteurs purement pécuniaires à expliquer pleinement les choix éducatifs et professionnels observés (Cunha and Heckman, 2007; D'Haultfoeuille and Maurel, 2013; Arcidiacono et al., 2020). Dans ce chapitre, j'apporte un nouvel éclairage sur cette question clé en comparant et en quantifiant les rôles des espoirs de gains et des facteurs non pécuniaires dans la décision de fréquenter l'université au Royaume-Uni. Des travaux récents ont exploré les autres facteurs au-delà du salaire pour expliquer es choix éducatifs et professionnels, principalement aux Etats-Unis : (Cunha and Heckman, 2007; Arcidiacono et al., 2020). Ces articles utilisent un terme résiduel pour capturer les facteurs non pécuniaires - faute de mesure directe - et constatent qu'ils jouent un rôle majeur dans les décisions en matière d'éducation et de choix de profession. Les travaux de Boneva and Rauh (2020) constituent une exception : ils ont réalisé une enquête auprès d'étudiants de l'enseignement secondaire au Royaume-Uni, en incorporant des attentes concernant les facteurs pécuniaires et non pécuniaires.

Ma contribution s'appuie sur ce travail, en exploitant un jeu de des données important sur les résultats observés et les attentes des jeunes concernant les facteurs non pécuniaires. Je propose un modèle pour le choix de l'université en utilisant les données d'un échantillon représentatif, qui contient les attentes des jeunes pour leur avenir et les résultats réalisés. J'utilise mon modèle pour étudier les facteurs qui affectent la fréquentation de l'université en Angleterre, en répondant à trois questions clés : (i) Quelle est l'importance des attentes en matière de revenus par rapport aux autres facteurs pour les jeunes de 16 à 18 ans lorsqu'ils décident d'aller à l'université ? (ii) Quels sont les facteurs à l'origine de l'écart des niveaux d'études entre les étudiants potentiels favorisés et moins favorisés ? (iii) Comment l'importance de ces facteurs dans leur décision a-t-elle changé entre les années 1980 et aujourd'hui ?

Je constate que les facteurs non liés aux revenus jouent un rôle encore plus important que dans les travaux précédents, les facteurs non pécuniaires étant quatre fois plus importants que les revenus espérés dans la décision de fréquenter l'université. Ce résultat souligne la diversité des coûts et des avantages que les jeunes prennent en compte lorsqu'ils décident de faire des études. Je constate également que les attentes en matière de revenus sont similaires pour tous les groupes socio-économiques, ce qui suggère que

les différences pour d'autres facteurs sont entièrement responsables de l'écart observé en matière de niveau d'études.

L'écart actuel entre les jeunes issus de milieux favorisés et ceux issus de milieux moins favorisés n'est pas dû aux différences de revenus attendus, ni aux difficultés à obtenir des financements. Pour remédier à ce déséquilibre socio-économique, les décideurs politiques devraient se concentrer sur d'autres aspects de la vie universitaire, des aspects plus faciles à influencer que les revenus attendus et moins coûteux que la réduction des frais d'inscription. J'utilise des informations détaillées sur les attentes des jeunes concernant la vie à l'université et après l'université pour décomposer les *autres facteurs* en catégories plus significatives et plus pertinentes pour les politiques. En séparant les attentes concernant l'endettement et les coûts liés à la fréquentation de l'université des autres facteurs, je constate que les facteurs financiers ne jouent pas un rôle majeur dans leur décision. Au contraire, il semble que les jeunes qui s'inquiètent le plus de l'impact de la dette des prêts étudiants - et des autres coûts liés à la fréquentation de l'université - sont ceux qui sont les plus susceptibles de fréquenter l'université. Enfin, le bénéfice pécuniaire d'un diplôme universitaire semble avoir peu changé au cours des dernières décennies, l'augmentation de la fréquentation étant due à de meilleures attentes concernant les facteurs non pécuniaires associés à un diplôme universitaire.

Chapitre 2 — Le rendement salarial de l'enseignement supérieur selon les compétences et les aptitudes

Alors que le premier chapitre s'intéressait aux rendements *ex ante* (attendus) de l'université, le deuxième chapitre estime les rendements *ex post* (réalisés) liés à l'obtention d'un diplôme. Cette étude s'inscrit dans le cadre d'une vaste littérature sur les rendements de l'éducation, depuis les effets (moyens) d'une année de scolarité supplémentaire jusqu'à des analyses plus précises des effets du franchissement de jalons éducatifs clés, c'est-à-dire l'obtention d'un diplôme d'études secondaires ou universitaires. Une difficulté majeure a été de séparer les effets de la capacité préalable de ceux de l'éducation, une question rendue encore plus problématique par des preuves récentes de la grande hétérogénéité des rendements de l'éducation. James Heckman et ses coauteurs ont été les premiers à mettre au point une méthodologie permettant de résoudre cette difficulté, en utilisant des mesures (bruitées) de la capacité antérieure ainsi que des résultats ultérieurs observés pour capturer (et contrôler pour) la capacité (non observée) et l'hétérogénéité des rendements de l'éducation. Une littérature différente mais connexe a souligné l'importance des compétences cognitives et non cognitives pour les réussites scolaire et ultérieure. Heckman et ses coauteurs ont étendu leur méthode pour permettre l'utilisation de multiples composantes des aptitudes, mais comme ils s'appuient sur des modèles factoriels, celles-ci doivent entrer de manière linéaire dans les fonctions de salaire

ou de production, ce qui interdit toute interaction entre les aptitudes cognitives et non cognitives.

Je développe une nouvelle méthode pour estimer les rendements des investissements en capital humain en fonction des capacités cognitives et non cognitives antérieures d'un individu, en permettant les interactions entre les différentes composantes des capacités, et j'utilise ce cadre pour estimer les rendements salariaux d'un diplôme universitaire en Angleterre. Je m'appuie sur des travaux récents qui étudient la formation du capital humain (voir Cunha et al. 2006 pour une revue), et j'incorpore des idées issues de l'abondante littérature sur les rendements de l'éducation (voir Card 2001 pour une revue). J'utilise ensuite ce cadre pour répondre à deux questions clés : comment séparer les effets de la capacité (cognitive) des effets de l'enseignement supérieur ? Les étudiants les plus brillants fréquentant les meilleures universités, ces dernières ajoutent-elles une prime importante aux salaires que ces étudiants peuvent espérer obtenir, ou ces étudiants très doués auraient-ils reçu des salaires plus élevés même sans diplôme ?

Pour répondre à ces questions, une contribution méthodologique clé a été développée. Je développe l'identification non paramétrique, et une stratégie d'estimation parcimonieuse, de deux objets clés (et connexes) : (1) les capacités cognitives et non cognitives d'un jeune au moment où il entre à l'université ; (2) les rendements (salariaux) du diplôme universitaire en fonction de ces capacités à l'entrée dans le cycle universitaire. En utilisant les récentes avancées dans l'identification des modèles de mélange discret (Bonhomme et al., 2017), ma stratégie d'identification nécessite moins d'observations et permet une forme fonctionnelle plus flexible que les principales approches actuelles de la littérature (voir par exemple Cunha et al., 2010). Je réalise cela dans un cadre traçable en supposant que la distribution du capital humain a un support discret, et que les individus ayant le même niveau de capital humain sont de même type latent. L'identification non paramétrique garantit que je n'ai pas besoin de m'appuyer sur des hypothèses paramétriques ou de forme fonctionnelle difficiles à valider. L'identification nécessite une mesure supplémentaire (brute) ou un instrument, une variable qui mesure ou affecte l'investissement en capital humain (c'est-à-dire la fréquentation de l'université), mais qui est indépendante des salaires, conditionnée par le capital humain avant l'investissement. Ce cadre permet de prendre en compte des effets non-linéaires dans les rendements de l'enseignement supérieur en fonction des capacités cognitives et non cognitives qui ne sont pas prises en compte par les principales approches actuelles de la littérature. Je montre empiriquement que ces effets non non-linéaires sont importantes dans à le cas de l'enseignement supérieur en Angleterre.

J'applique ce cadre aux données de la British Cohort Study, un ensemble de données longitudinales contenant des informations sur 16 000 personnes nés la même semaine en avril 1970. Les résultats suggèrent des bénéfices importants de l'université pour ceux qui ont une faible capacité à l'entrée : ils " rattrapent " en termes de salaire

leurs pairs non diplômés de l'université qui possédaient des capacités beaucoup plus élevées avant l'université. Cependant, le rendement salarial lié à la fréquentation de l'université est généralement croissant en fonction du capital humain antérieur, ce qui signifie qu'un diplôme universitaire accroît l'inégalité salariale entre les diplômés. Ce résultat n'est pas surprenant à la lumière des recherches sur la formation du capital humain pendant l'enfance, qui montrent que "les compétences engendrent les compétences, [et] l'apprentissage engendre l'apprentissage" (Cunha et al., 2006, p. 799), c'est-à-dire que le capital humain préexistant augmente l'efficacité des investissements ultérieurs en capital humain. Je prouve également l'existence de non-linéarités dans les rendements de l'enseignement supérieur, en particulier dans les interactions entre les compétences cognitives et non cognitives. Il n'est pas possible de prendre en compte ces interactions en utilisant les approches habituelles d'estimation qui reposent sur un modèle additif pour l'identification et l'estimation.

Chapitre 3 — Une approche non paramétrique par mélange fini de l'estimation par différence de différence, avec une application à la formation en cours d'emploi et aux salaires

avec Robert GARY-BOBO, Julie PERNAUDET, et Jean-Marc ROBIN.

Le troisième chapitre s'écarte de l'étude de l'enseignement supérieur pour se concentrer sur un autre investissement clé dans le capital humain : la formation professionnelle. La formation est depuis longtemps au centre des préoccupations des responsables politiques qui cherchent à faciliter le réemploi des travailleurs des industries perturbées par l'automatisation et le commerce international. Elle apparaît également comme un substitut naturel à l'enseignement supérieur, ce qui peut contribuer à atténuer les effets de l'enseignement supérieur sur l'inégalité des salaires mis en évidence au chapitre 2. Cependant, les preuves du rendement salarial de l'enseignement supérieur sont mitigées, la plupart des auteurs constatant de faibles effets positifs, bien que des analyses plus récentes aient trouvé des effets allant jusqu'à 10 % d'augmentation de salaire. Cependant, le recours à la formation professionnelle est généralement faible, ne concernant que 40 % des travailleurs en France et seulement 20 % aux Pays-Bas. Des rendements aussi élevés de la formation sont en contradiction avec l'impact perçu de la formation. La sélection, l'hétérogénéité non observée et le traitement endogène pourraient-ils entraîner des biais si importants que les méthodes d'évaluation statistique standard (variables instrumentales, modèles de sélection, etc.) sont incapables de fournir des estimations fiables ? Ou bien la perception est-elle tout simplement fausse ? Notre article apporte des contributions originales à cette question, tant sur le plan empirique que méthodologique.

Nous développons et appliquons une méthodologie originale à de nouvelles données administratives et d'enquêtes salariés-employeurs couplées afin de mesurer le rendement

salarial de la formation en France. Notre approche s'inscrit dans l'esprit de l'estimation par différence de différence, mais nous utilisons une combinaison de restrictions d'exclusion motivées économiquement et de mélanges discrets (pour capturer l'hétérogénéité non observée) afin d'assouplir l'hypothèse de tendances communes habituellement requise dans de telles analyses. Nous prouvons l'identification non-paramétrique de notre modèle et démontrons une stratégie d'estimation viable via l'algorithme expectation-maximisation (EM).

Empiriquement, nous trouvons de effets moyens limités de la formation sur les salaires d'environ 1% . Notre cadre permet aux rendements de varier selon les *types*,[1] et nous constatons une hétérogénéité significative des effets de la formation entre ces différents types. Pour certains types, nous estimons des effets de traitement (i.e. entrainement) de plus de 10 %, tandis que pour d'autres, les effets de la formation sur les salaires sont légèrement négatifs. Nous sommes en train d'étendre ce travail pour essayer de mieux comprendre les facteurs de ces différences dans les effets de la formation. Ces résultats sont importants étant donné l'accent mis par les gouvernements du monde entier sur la formation comme solution aux changements du marché du travail, notamment ceux induits par la technologie. Si les avantages de la formation diffèrent d'un travailleur à l'autre, alors des politiques mal informées en matière de formation des travailleurs pourraient être inefficaces et même conduire à une augmentation des inégalités salariales. Une autre conséquence politique directe de notre travail est la mise en évidence de l'importance de la fourniture d'informations sur les possibilités de formation aux travailleurs par leurs employeurs.

Bibliography

Arcidiacono, P., V. J. Hotz, A. Maurel, and T. Romano (2020): "Ex Ante Returns and Occupational Choice," *Journal of Political Economy*, 128, 4475–4522.

Boneva, T. and C. Rauh (2020): "Socio-Economic Gaps in University Enrollment: The Role of Perceived Pecuniary and Non-Pecuniary Returns," *HCEO Working Paper*, 83.

Bonhomme, S., K. Jochmans, and J.-M. Robin (2017): "Nonparametric estimation of non-exchangeable latent-variable models," *Journal of Econometrics*, 201, 237–248.

Card, D. (2001): "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69, 1127–1160, publisher: [Wiley, Econometric Society].

[1]Nous utilisons des types latents discrets pour capturer l'hétérogénéité non observée.

Cunha, F. and J. J. Heckman (2007): "Identifying and Estimating the Distributions of Ex Post and Ex Ante Returns to Schooling," *Labour Economics*, 14, 870–893.

Cunha, F., J. J. Heckman, L. Lochner, and D. V. Masterov (2006): "Interpreting the Evidence on Life Cycle Skill Formation," in *Handbook of the Economics of Education*, Elsevier, vol. 1, 697–812.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010): "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 78, 883–931, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA6551.

D'Haultfoeuille, X. and A. Maurel (2013): "Inference on an extended Roy model, with an application to schooling decisions in France," *Journal of Econometrics*, 174, 95–106.